Driving while Black (and male, and young, and...): Evidence of disparities at the margin and the intersection

Frank R. Baumgartner, Leah Christiani, Derek Epp,
Santiago Olivella, Kevin Roach, & Kelsey Shoub[1]

August 23, 2018

Abstract

A police officer must make a quick judgment about whether to search a vehicle following a routine traffic stop. Similarly, officers must decide whether to carry out an arrest if the conducted search is successful. Informing profiles of whom to investigate, the psychology literature on perceptions of risk tells us that those perceived more threatening will face greater suspicion, and accordingly a greater likelihood of search and arrest. In the context of policing, the intersection of age, gender, and race are easily visible cues that officers may use, and indeed we know that young men of color are more likely to observe adverse police outcomes than older white women. When assessing whether disparities exist across intersectional identity groups in search and arrest decisions, common estimation strategies fall prey to the problem of infra-marginality: absent systematic disparities, the rates at which minority and white drivers are searched and arrested might still be different if the groups have different risk distributions. To avoid these pitfalls, we rely on a hierarchical Bayesian latent thresholds model developed by Simoiu, Corbett-Davies, and Goel 2017 and on a covariate-balancing matching algorithm developed by Imai and Ratkovic (2014). Using these techniques, this paper identifies important disparities in the accuracy of police inference about criminal suspicion, the rates at which drivers are subjected to fruitless and unwarranted search, and who is arrested even after being discovered with contraband of various amounts.

Paper prepared for presentation at the annual meeting of the American Political Science Association, Boston, MA, August 30-September 2, 2018.

[1]Frank R. Baumgartner is the Richard J. Richardson Distinguished Professor of Political Science at the University of North Carolina at Chapel Hill. Santiago Olivella is an assistant professor at the University of North Carolina at Chapel Hill. Derek Epp is an assistant professor at the University of Texas at Austin. Kelsey Shoub is a post doctoral research associate at the University of Virginia. Leah Christiani and Kevin Roach are graduate students at the University of North Carolina at Chapel Hill.

1 The problems of infra-marginality and intersectionality

When a police officer stops a vehicle, they must make a quick judgment about whether the driver and/or vehicle is suspicious enough to warrant a search. In doing so, the officer is effectively estimating the likelihood that they may find contraband based on a number of data inputs available to her at the time — including the driver's criminal record, their behavior both prior and during the stop, and the state of their cars. Whatever those inputs may be, if the likelihood is high enough, the officer can be expected to conduct a search. When officers use different estimated likelihood levels to search members of different groups, it is reasonable to say that the decision process is unfair to the group for which a stricter a standard is being used.

Prior studies on police traffic stops have analyzed the process in two broad ways. First, some have employed a bench-marking method, which simply compares search rates, or other measures of disparate treatment, across identity groups. When minority drivers are searched more than whites, even after controlling for multiple factors, these disparities are said to be unwarranted.

This method, however, does not account for whether this increased scrutiny may be warranted. Minority drivers may be searched more, but they may also be found to be committing more arrest-able or search-worthy offenses on average. To address this problem, Becker (1957, 1993) first proposed outcome tests. The outcome test compares the success rate of some decision. In traffic stops, this is often the comparison of hit rates, or how often a search results in contraband being found. When there are heightened levels of searches among minority groups, but simultaneously lower levels of contraband found, then, according to Becker, discrimination is said to exist. Indeed, studies have found that minority drivers are more likely to endure searches following a stop while simultaneously making up the bulk of fruitless searches, indicating that these higher search rates do not yield more contraband hits (Baumgartner et al., 2017; Baumgartner and Shoub, 2018).

However, even when we observe differential levels of police targeting of certain identity groups, this targeting may not necessarily be unwarranted. If a group is more likely to be breaking the law, carrying contraband, or engaging in an arrest-worthy offense, then a higher degree of targeting by the police may well be justified. Similarly, a different success rate of searches across targeted groups could be attributable to very different underlying distributions of the probability that any one member of the group is, in fact, carrying contraband. In neither case is the officer using

a stricter standard to conduct searches across certain groups; it is the relative frequency with which individuals from those groups can be expected to be guilty that is different. Similar arguments can be made to explain observed differences in arrest rates.

Imagine that it is, for some reason, harder to tell whether members of one group are hiding contraband than to see it in another group, based on the information available to an officer before conducting a search. Further, imagine that the same share of both groups are indeed concealing it. If officers are more accurate in their searches of one group than the other, because it is easier to discern in that group, their contraband hit rates for that group will be higher. This would be true even if the underlying rates of criminality do not differ. The raw hit-rate analysis proposed by Becker assumes that the underlying distributions are similar; they may not be.

In general, and in order to assess whether there are systematic disparities in the way officers decide to conduct searches or arrest individuals, we need information about the likelihood of carrying contraband *at the margin* — that is, we should compare people who almost were not searched or arrested across groups. Then we should evaluate whether their likelihood of being guilty is different enough to suggest that, in fact, different standards were used in the decision. The problem, of course, is that we only observe *whether* a driver is searched or arrested, which gives us no indication of whether the driver was marginal in the sense that they barely made the cut of what constitutes a justifiable search or a reasonable arrest. This problem, identified by Ayres (2002) as the "problem of infra-marginality" (i.e. the problem of not having access to information about the likelihood of carrying contraband at levels of aggregation *below* the margin), prompted Simoiu et al. (2017) to define a test of discrimination that helps us get around this issue in the context of search decisions.

Their proposed "threshold test" (and the further refinements made to it by Pierson, Corbett-Davies and Goel, 2017) jointly estimates decision thresholds (i.e. the level at which the officer decides to conduct a search) *and* risk distributions (i.e. distribution over probabilities that a member of an group is carrying contraband). By doing so, the model is able to identify information at the right level of aggregation, thus enabling comparisons *at the margin*. Employing their proposed test, we evaluate whether disparities remain in traffic stops after appropriately estimating risk distributions by identity group.

In this paper, we test for the existence of identity-based disparities in traffic stop outcomes at multiple levels and with multiple methods. We first demonstrate significant disparities at the

initial stage of a traffic stop by analyzing the officer's decision about whether or not to search a car. Here, we use the threshold test to estimate both the officer's decision threshold and the driver's risk distribution. Then, we analyze the arrest stage. We compare similarly situated drivers who were already searched to see who is arrested following the discovery of a certain amount of contraband, given otherwise identical circumstances (e.g., same time of day, same police agency, same age group, gender, etc.). We find that even when contraband levels are equal, black and Hispanic male drivers experience a greater rate of arrest, compared to whites.

Further, we expand our understanding of identity-based disparities in traffic stop outcomes by considering the intersection of multiple identities. As we discussed earlier, officers take in and evaluate the entire context of the stop when estimating the likelihood that the driver is carrying contraband. They may consider the time of night, part of town, style of car, or how the driver is acting. At least partially, they may also be relying on some widely held stereotypes about identity groups, even if subconsciously.

Perceptions that arise from such stereotyping operate on multiple levels simultaneously. So if search and arrest thresholds are being estimated with the use of identity-based stereotyping in any way, we would expect that this stereotyping would operate not only by race, but by multiple demographic characteristics simultaneously (Fiske, 1993). This is especially true for the most readily available and visible markers of identity like race (Devine, 1989; Devine and Elliot, 1995), age, and gender (Brewer and Lui, 1989).

In this study, we focus on race and gender, and expect that search thresholds are likely to differ based on the driver's perceived identity. Specifically, the intersection of black or Hispanic (race) and male (gender) is likely to produce the greatest degree of suspicion (i.e. lowest search threshold) because this combination is perhaps the most recognized stereotype of criminality (Baumgartner et al., 2017; Baumgartner and Shoub, 2018; Welch, 2007). Previously, studies were not able to estimate the risk distribution that these groups may simply be more likely to carry contraband, and thus were left open to the critique that this targeting may be justified. Now, though, we are able to perform a more robust analysis of the way that search and arrest thresholds differ by race and gender. We expect that first, disparities will persist even after accounting for identity-based risk distributions with the threshold test, and second, that disparities affecting black and Hispanic male drivers (compared to their white counterparts) will be larger than the those affecting black

and Hispanic female drivers (compared to their white counterparts).

2 Disparities in search decisions

The model

In general, the model proposed by Pierson, Corbett-Davies and Goel (2017) (or PCG; which in turn is based on Simoiu et al., 2017) relies on the idea that a stopped individual can belong to one of two classes, depending on whether they are carrying contraband or not. Let Y_{gd} be a random variable denoting class membership, with $Y_{gd} = 1$ indicating that the member of group g policed by agency d is carrying contraband, and let π_{gd} denote the probability that a randomly chosen individual in said group is guilty of carrying contraband.

Although police officers cannot directly observe Y_{gd}, members of each class are expected to emit observable *signals* — an abstract construct encompassing the collection of factors routinely available to officers during a stop, such as behavioral indicators or an individual's criminal record. Crucially, these signals are believed to be produced by *different* distributions depending on an individual's class membership, so that guilty individuals produce higher signals than innocent ones on average. Letting $X_{gd} = x$ denote the signal emitted, we can thus define random variable $P_{gd} = \Pr(Y_{gd} = 1|X_{gd})$ — the probability that an individual is guilty of carrying contraband given the signal they emit. A distribution over perceived "risks" $p_{gd} = \Pr(Y_{gd} = 1|X_{gd} = x)$, which (given an estimate of $\pi_g d$) can be evaluated by the stopping officer, captures heterogeneity in the extent to which stopped members of a given identity group are deemed suspicious.

The model then introduces a *suspicion threshold* t_{gd} above which an officer is assumed to always decide to conduct a search. Defined in the scale of risks P_{gd}, this threshold of suspicion can be interpreted as the conditional guilt probability (given an observed signal) above which a search is deemed to be justified. If these thresholds are systematically lower for some groups, then we will conclude that searches were typically triggered under less suspicious circumstances for members of these groups, and that therefore they were discriminated against. In other words, finding that there are discernibly different thresholds across groups within the same agency would indicate that different standards of suspicion are being used to search members of different identity groups, and those for which the standard is lower could be thought of as being the target of an unfair decision

process. This is what Simoiu et al. (2017) call *the threshold test*. Our inferential goal will thus consist of learning about these thresholds using data on search decisions and their outcomes, and to explore the implications regarding differential enforcement of search standards across identity groups.

To do so, we rely on the estimation strategy proposed by PCG. They begin by defining the overall proportion of searched individuals as $\lambda^s_{gd} = \Pr(P_{gd} > t_{gd})$ (i.e. the probability that a randomly selected member of g stopped by an officer in d exceeds t_{gd}), and the proportion of successful searches (or *hit rate*) as $\lambda^h_{gd} = \mathbb{E}[P_{gd}|P_{gd} > t_{gd}]$ (i.e. the expected probability that a member of g searched by an officer in d is actually carrying contraband).

For each stop i, they then define two observed random variables: $S_{g[i],d[i]}$, equal to 1 if a search occurs (where $g[i]$ denotes the group associated with stop i, and similarly for $d[i]$), and $H_{g[i],d[i]}$, equal to 1 if a search results in a "hit" (i.e. if a search yields contraband). For each individual stop, their model defines the following data-generating process of searches and hits:

1. The officer observes a risk $p_{g[i],d[i]}$, and compares it to the group and agency-specific threshold $t_{g[i],d[i]}$.

2. If $p_{g[i],d[i]} > t_{g[i],d[i]}$, then a search occurs (i.e. $S_{g[i],d[i]} = 1$); otherwise, no search occurs. Thus, $S_{g[i],d[i]} \sim \text{Bern}\left(\lambda^s_{g[i],d[i]}\right)$

3. If a search occurs, then the officer finds contraband (i.e. $H_{g[i],d[i]} = 1$) with probability $p_{g[i],d[i]}$. Otherwise, $H_{g[i],d[i]} = 0$. Formally, $H_{g[i],d[i]} \sim \text{Bern}\left(\lambda^h_{g[i],d[i]}\right)$.

For a given number of independent stops n_{gd}, we can thus model the total numbers of searches ($S_{gd} = \sum_{i=1}^{n_{gd}} S_{g[i],d[i]}$) and successful hits among those searched ($H_{gd} = \sum_{i=1}^{S_{gd}} H_{g[i],d[i]}$) as draws from binomial distributions with probability parameters given by λ^s_{gd} and λ^h_{gd}, respectively.

Computing λ^s_{gd} and λ^h_{gd} requires the definition of a distribution for the risk P_{gd}. Although commonly used families like the Beta and logistic-Normal have the right support and could therefore be used to model the risk, PCG show that using an expressive family of distributions — which they call *discriminant distributions* — can lead to dramatic speed-ups in the inferential task. Closely related to homoskedastic linear discriminants, these risk distributions assume that signals are Gaussian with a common variance, and are fully parameterized by the proportion of guilty individuals

π_{gd} and the difference between guilty and innocent signal distribution means δ_{gd}. Complementing the threshold of suspicion, the risk distribution can also shed light on other ways in which particular groups are discriminated against. For instance, we can use them to estimate the extent to which innocent members of different groups are wrongfully targeted for a search (i.e. the false positive rates), or the extent to which officers are in a position to discern accurately between guilty and non-guilty individuals from different groups (i.e. the classification accuracy across these groups).

Finally, the model proposed by PCG addresses an identification problem that hinders the simultaneous estimation of thresholds and risk distributions. In general, threshold values and risk distributions that would result in very different assessments of disparity could be consistent with the same observed search and hit rates, thus making it impossible to estimate both the suspicion thresholds and the risk distribution parameters from these data without making additional assumptions. Their solution, which is more thoroughly discussed in Simoiu et al. (2017), requires assuming that the risk distribution parameters decompose into group- and agency-specific terms drawn from common distributions, so that

$$\begin{aligned}\pi_{gd} &= \text{logit}^{-1}(\pi_g + \pi_d) \\ \pi_g &\sim N(\mu_g, 1) \\ \pi_d &\sim N(0, 1)\end{aligned}$$

and similarly for δ_{gd}. The solution has the added benefit of allowing for partial pooling across identity groups and police agencies, thus producing more precise parameter estimates in instances in which few stops are observed. The Bayesian model is completed by defining prior distributions for all remaining model parameters.[2]

Results

To estimate the parameters and derived quantities of interest, we rely on data on traffic stops conducted by police departments in North Carolina between 2002 and 2016. The data amounts to almost 9,400,000 stop instances,[3] which are coded in terms of whether a search was conducted and

[2] See Pierson, Corbett-Davies and Goel (2017) for details.

[3] We exclude agencies for which fewer than 200 stops were conducted in the 15 years of study. Data from the State Highway Patrol were excluded because they typically do not include the time of day, which is an important control

whether, upon searching the suspect, a meaningful amount of illegal contraband was discovered.[4] For each stop instance, we take note of the suspect's race ('black', 'white', 'Hispanic', or 'Other') and gender (Female or Male), and define identity groups based on combinations of these attributes. The posterior distribution over parameters of interest is explored by obtaining 2500 samples (after burn-in) via HMC with 5 Markov chains, simulated using Stan.[5]

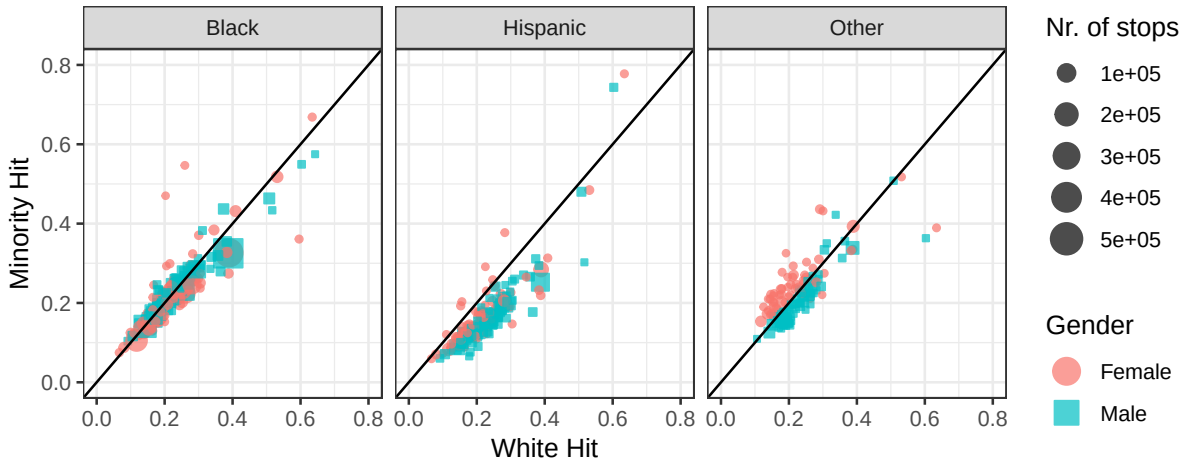

Figure 1: **Comparison of white and minority search success rates by race and gender**: The figure depicts the estimated success (or "hit") rates of searches (i.e. the probability of finding contraband once a search is initiated) for whites (x-axis) against the estimated hit rate for black (left panel), Hispanic (center) and Other minorities (right), for both Women (salmon circles) and Men (green squares). Points below the 45° line would indicate that, when the minority is searched, the probability of finding contraband is *lower* than it is for whites, and is therefore evidence of non-random disparities affecting the minority. Points are scaled by the number of stops experienced by the corresponding minority over a 15 year period.

Originally proposed by Becker (2010), the *outcome test* of racial disparities in search decisions compares the success rates of searches among whites to those of racial minorities. Even if officers expect different racial groups to have different probabilities of carrying contraband, Becker suggests that the success rate of searches (i.e. the probability of finding contraband, conditional on a search being conducted) should be comparable across identity groups if officers are deciding

variable for us. The data then represent virtually every sheriff and police department in the state.

[4]To reduce concerns about inconsequential quantities of contraband being coded as "hits," we define a successful contraband discovery as an instance in which the amount found is higher than the 75th percentile of observed quantities for the type of contraband recorded. Although all our results are robust to the more standard definition of a successful "hit," we believe that our operational definition is more consistent with the idea of a meaningful criminal condition being discovered.

[5]We follow Pierson, Corbett-Davies and Goel (2017) on the definition of hyperprior parameters. Convergence is assessed by visual inspection of the chains' traceplots, as well as by $\hat{R}$ values well that are well below 1.1 for all model parameters.

to carry out searches based on the same levels of evidence/suspicion. A lower success rate for one group relative to another would suggest that a different (viz. higher) standard of suspicion is being applied to the former than to the latter, and thus that a lopsided decision process is taking place.

Figure 1 shows the estimated "hit" (or success) rates among white suspects against the hit rates among racial minorities, for both men and women. Points below the diagonal in each panel are indicative of the sort of biased decision-making that the outcome test was designed to identify. Overall, the outcome test suggests that, while white and black drivers are found to be carrying contraband at roughly similar rates (i.e. their hit rates are typically close to the diagonal on the left panel of Figure 1), most Hispanic suspects and some suspects in other racial categories appear to be over-policed, with hit rates that are systematically lower than those of white suspects stopped by officers in the same agencies. And when such disparities are apparent, men in particular (depicted as turquoise squares in Figure 1) seem to be the likely targets of these searches.

While suggestive, the outcome test is not without limitations. As we discussed earlier, the problem of infra-marginality can result in hit rates that are similar under scenarios in which the decision to conduct a search made use of different standards across groups. Similarly, even when hit rates look dissimilar across groups, the result could be an artifact of differences in the underlying relative frequency with which the over-policed group is carrying contraband rather than of a double-standard on the part of the officers. Accordingly, a better measure of systematic discrepancies is given by a comparison of the *thresholds* of suspicion that are estimated in the model of Pierson, Corbett-Davies and Goel (2017).

Figure 2 shows the estimated thresholds of suspicion for white vs. black (left), Hispanic (center) and Other (right) racial minorities, once again divided by gender (with men represented as turquoise squares). Points below the diagonal indicate that, for a given agency, the estimated level of suspicion needed to prompt a search of a white driver is higher than the level needed to justify a search of the minority driver — a direct comparison of individuals "at the margin" that suggests minority drivers are more likely to be searched under less suspicious circumstances. The results depicted in Figure 2, which indicate that most agencies tend to use lower levels of suspicion to prompt minority searches, are consistent with the idea that both black and Hispanic drivers are typically searched when there appears to be less cause for it.

Figure 2 also reveals that disparities in the levels of suspicion that prompt searches are

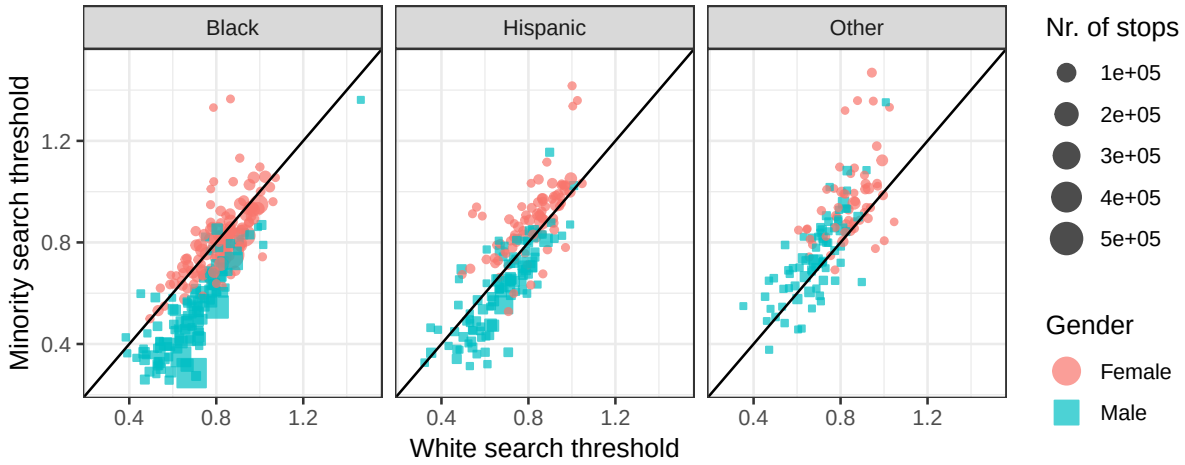

Figure 2: **Comparison of white and minority thresholds of suspicion by race and gender**: The figure depicts the estimated threshold of suspicion used by agencies for whites (x-axis) against the estimated threshold of suspicion used for black (left panel), Hispanic (center) and Other minorities (right), for both Women (salmon circles) and Men (green squares). Points below the 45° line would indicate that the minority is searched under *less* suspicious circumstances than whites, and is therefore evidence of disparities affecting the minority. Points are scaled by the number of stops experienced by the corresponding minority over a 15 year period.

most apparent among men. In fact, although black women can be expected to be searched under similar levels of suspicion as their white counterparts, black men are estimated to be subject to lower thresholds in all but a few of the agencies in our data set. The same appears to be true, albeit to a lesser extent, when we consider Hispanic men and women. In turn, women in Other racial categories appear to be given far more latitude than their white counterparts, as the opposite pattern is apparent for them: in most agencies in our sample, women in these other racial categories are only searched when they appear more suspicious than their white counterparts.

Are the searches prompted by different levels of suspicion justified? Perhaps black and Hispanic men are believed to be better are concealing their illicit activities, so using a lower threshold of suspicion is justified. Although the comparison of hit rates among whites and Hispanic drivers presented in the central panel of Figure 1 would certainly serve as evidence against this notion, the fact that black and white drivers seem to be found to be carrying meaningful amounts of contraband at similar rates could be used to justify this alternative view of the use of different thresholds of suspicion. The problem with such a justification, however, is that it remains blind to the additional costs imposed on innocent black drivers who are subject to a search but are found to be carrying no contraband.

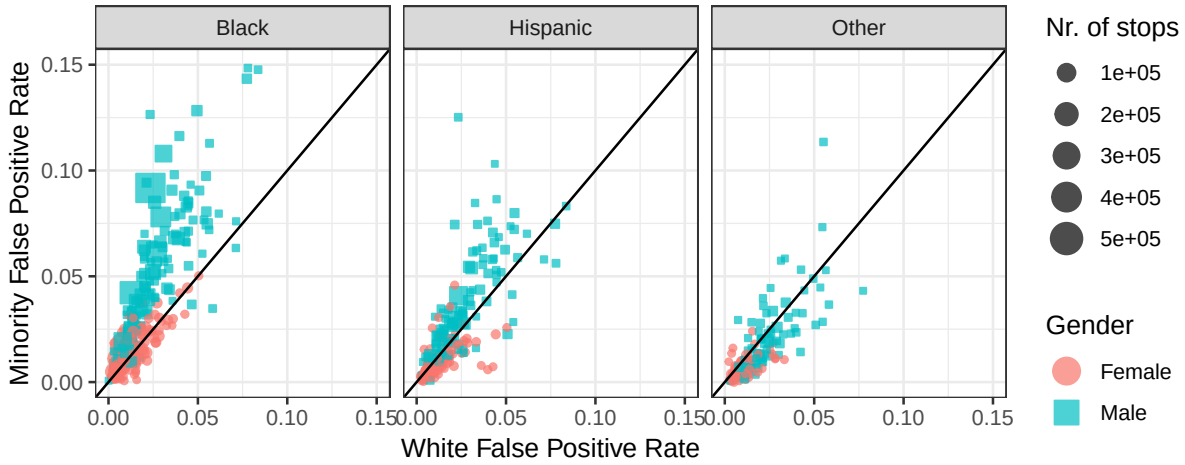

Figure 3: **Comparison of white and minority false-positive rates by race and gender**: The figure depicts the estimated proportion of innocent people who are fruitlessly searched (i.e. the false positive rate) of agencies for whites (x-axis) against the estimated false positive rate for black (left panel), Hispanic (center) and Other minorities (right), for both Women (salmon circles) and Men (green squares). Points above the 45° line indicate that a higher proportion of innocents in the minority are subjected to unjustified searches, and is therefore evidence of disparities affecting the minority. Points are scaled by the number of stops experienced by the corresponding minority over a 15 year period.

Figure 3 presents a comparison of estimated *false positive rates* — the proportion of innocent members of an identity group that are wrongfully subjected to a search, also known as the proportion of Type I errors. In general, if decisions to search are being made in a way that equally protects the rights of all innocent drivers, then false positives should occur at roughly similar rates across identity groups, and points should fall close to the diagonals in all panels of Figure 3. What the data and model suggest, however, is that innocent black drivers are much more likely to be targeted for a fruitless search. This is evidenced by the fact that, in most agencies, the estimated false positive rates tend to be higher for members of these identity groups than for their white counterparts (i.e. most points fall *above* the diagonals in the left and central panels of Figure 3). This over-targeting of innocent black men may well be the source of historical alienation and anti-force sentiment among members of these populations (Lerman and Weaver, 2014).

Effectively, these fruitless search decisions are the result of a failed classification exercise on the part of the officer: although the searched drivers were classified as likely members of the guilty class, the fact that they were found to be carrying no contraband revealed their true membership in the innocent class. Accordingly, we can evaluate just how much better or worse officers in an agency

are at classifying drivers of a particular identity group *vis-à-vis* their white driver counterparts. Although there are many different ways of evaluating the quality of a classifier, the area under the receiver operating characteristic curve (or *AUCROC*) is a popular metric in the Political Science and statistical learning literatures (Friedman, Hastie and Tibshirani, 2001). The measure captures the probability that a driver chosen randomly from the guilty pool will have a perceived risk that is higher than that of a driver chosen randomly from the innocent pool. Thus, higher AUCROC values would indicate a more adequately discriminating classification exercise.

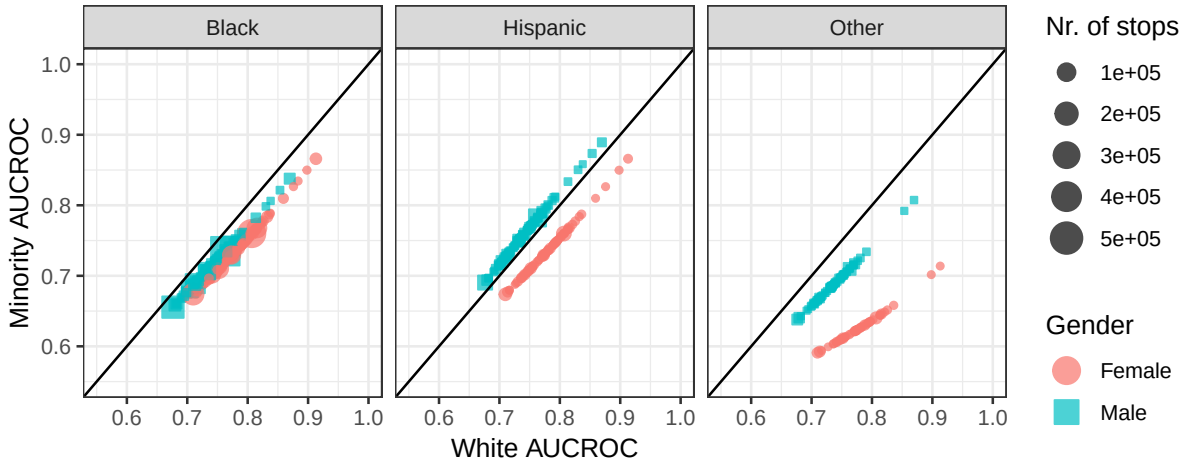

Figure 4: **Comparison of white and minority area under the receiver operating characteristic curve (AUCROC) by race and gender**: The figure depicts the agency-specific estimated classification accuracy into guilty and innocent classes, as measured by the AUCROC, for whites (x-axis) against the estimated AUCROC for black (left panel), Hispanic (center) and Other minorities (right), for both Women (salmon circles) and Men (green squares). Points below the 45° line indicate that whites are more accurately classified as guilty or innocent than their minority counterparts, and is therefore evidence of disparities affecting the minority. Points are scaled by the number of stops experienced by the corresponding minority over a 15 year period.

Figure 4 compares the estimated, agency-specific AUCROC values for men and women in racial minority groups against the AUCROC values for their white counterparts. Once again, points below the diagonal suggest officers in an agency are better at classifying white drivers than the corresponding minority drivers (i.e. they are better at correctly assessing whether a white driver is innocent or guilty). The figure shows that, for the most part, officers are better classifiers of white drivers. The one exception to this is the case of male Hispanic drivers, whom officers seem to be slightly better at classifying than their white counterparts.

Figure 4 also evidences a striking pattern when it comes to men and women in the Other

racial category. In general, it appears as though officers are particularly bad at sorting these types of drivers (and particularly women in these racial groups) into the guilty or innocent bins relative to their white counterparts. In their case, however, the misclassification is of a different kind altogether, and is driven primarily by Type II errors — a likely lack of searches of guilty individuals.

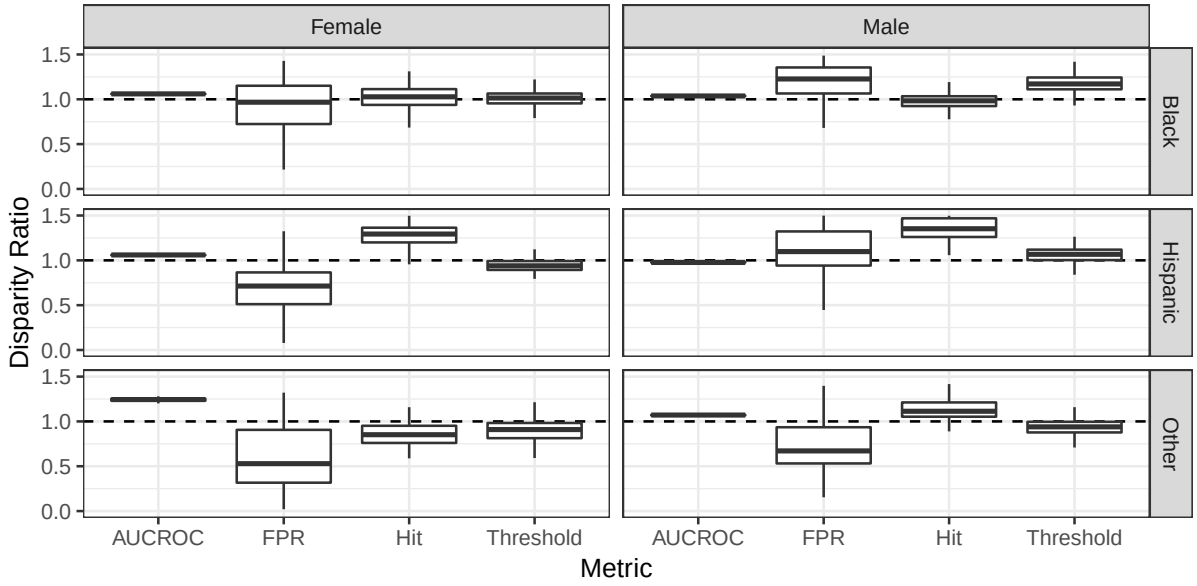

Figure 5: **Disparity ratios for different identity groups**: The figure shows the ratios of agency-level indicators for whites and members of different minorities. The ratios are computed so that values above one are indicative of a disparity that works against the minority.

To summarize these results, Figure 5 shows what we call the within-agency "disparity ratios" in search decisions, which evaluate the extent to which any of the indicators we have used (viz. the AUCROC, the false positive rate, the hit rate and the estimated threshold of suspicion) are suggestive of systematically unfair treatment of a minority group *vis-à-vis* white North Carolinians. For each indicator, we compute the ratio of values that would indicate the extent to which the minority is subjected to an unequal treatment when it comes to search decisions, so that values above 1 are suggestive of unequal treatment of the minority.[6]

The summary plot neatly presents our main findings: black and Hispanic men tend to be subjected to searches under consistently lower levels of suspicion. Among black men, the lower threshold of suspicion translates into almost 25% more wrongfully accused individuals when com-

[6]For most indicators, this means we take the ratio of its value as estimated for a minority identity group to the value estimated for the corresponding white group. In the case of the false positive ratio, higher values are indicative of worse outcomes. Accordingly, we reverse the ratio for the FPR indicator, in order to keep interpretation of the ratios consistent.

pared to their white counterparts. Among Hispanic men, this in turn translates into hit rates that are roughly 37% worse than they are for white men. Although black women tend to be treated roughly equally to their white counterparts, Latina women seem be searched under more lenient standards and are less likely to be wrongfully accused that white women. Finally, women in other racial groups appear to be the subject of the least stringent standards, with indicators that typically tip the scales in their favor *visà-vis* white women, resulting in patterns that are almost mirror images of those comparing black men to their white counterparts.

Do these disparities carry over to the next stage of the process, when officers must decide whether a successful hit leads to an arrest of the driver involved? We explore this issue using a different empirical approach that, once again, allows us to make inferences by comparing individuals at the margin that distinguishes those who were barely arrested to those who were arrested, but could just as easily have been allowed to leave.

3 Disparities in arrest decisions

When deciding to search or not to search, officers have a great deal of discretion: there are a range of circumstances that may count as probable cause, and even in the absence of probable cause an officer may request to conduct a consent search. In such environments—where officers have acknowledged discretion—it is reasonable to anticipate that minority and white drivers will be treated differently, as we find they indeed are. When discretion decreases, however, we might expect to see these disparities disappear. In the following section, we examine this question: following a successful search where contraband is found, do we still observe disparate treatment by a driver's race?

To evaluate this, we isolate male drivers who were searched and for whom contraband was found, and we then predict the probability of arrest for drivers from different races, given similar or identical contextual and demographic profiles (e.g., the drivers had similar amounts of contraband, were of similar ages, were arrested by an officer of the same police agency at the same time of day, for the same stop purpose, and so on). To better capture discrepancies across similar cases that result from differences in race only, we balance the data set on those contextual and demographic characteristics. Doing so allows us to establish whether the conditions under which male drivers from minority races are arrested differ systematically from those under which their white counterparts

are arrested.

Data & weights

To conduct our analysis of arrests at the margin, we rely on a slightly different set of observations than those used in our study of search discrepancies. First, for this part of the analysis, we focus on male drivers who were searched and for whom contraband was found. We do this to isolate a set of cases for which clear evidence of unequal treatment was found in our previous analysis. Additionally, we focus exclusively on three race-ethnicity groups: White non-Hispanic drivers, black non-Hispanic drivers, and Hispanic drivers. This is because we simply do not observe enough instances of male drivers of other races found with meaningful amounts of contraband to provide us with sufficient power to conduct statistical inference. As before, individual stops for which a search of a male driver is conducted and contraband found are aggregated by agency, age category, race, stop purpose, amount of contraband found, number of contraband types, and time of the stop. For each of the profiles defined by unique values of these variables, we then calculate the proportion arrested.[7] The proportion arrested can therefore be interpreted as the probability of being arrested for a given individual of that profile.

To calculate the proportion arrested by driver profile, the continuous variables must be collapsed into broader bins. After all, a 16 year-old does not differ that much from a 17 year-old nor does being stopped at 1:15 pm differ that much from being stopped at 2 pm. First, we collapse age into eight ten-year windows starting with those between 16 and 25 and ending with those between 86 and 95.[8] In the regression analysis that follows, those between 16 and 25 are the reference category. Second, time of day is record down to the second.[9] We collapse time of day into four categories: 11 pm until 5 am, 5 am until 11 am, 11 am until 5 pm, and 5 pm until 10 pm. In the regressions that follow, the 11 pm until 5 am time window is the reference category. Third, there are ten possible reasons for a stop recorded on the traffic stop form in North Carolina. As Epp, Maynard-Moody and Haider-Markel 2014 and Baumgartner and Shoub 2018 do in their respective

[7] We exclude incident-to-arrest searches, because the outcome has already been determined prior to the search.

[8] Those recorded as under 16 or over 95 are dropped from the analysis.

[9] A disproportionate number of stops are recorded as occurring exactly at midnight (0:00:00). As a result, we drop these observations from the analysis. This drops the entire State Highway Patrol and a small number of other observations. See Baumgartner and Shoub 2018, Appendix E.

books, we collapse these reasons into two categories: safety stops and investigatory stops.[10] Safety stops are those stops explicitly made for the protection of the general citizenry and to prevent unsafe driving, whereas investigatory stops are other stops made for the explicit or implicit purpose of an officer getting a look inside of a car.

Additionally, drivers may be found to carry varying amounts of contraband. When filling out the official form following a traffic stop, officers are first asked whether contraband is found, and then they are asked to indicate what amount of contraband is found. By indicating the amounts, they also indicate the type of contraband. If drugs are found, they are instructed to estimate the ounces, grams, dosages, or number pills found. If alcohol is found, they should estimate the pints and gallons present. If weapons are found, they simply enter the number found. If cash is found and seized, they should list a dollar amount. Finally, if other contraband is found or seized (ex. a stolen television), they are instructed to estimate the cash value. However, if values are less than one-half of a unit, the state's software system rounds these values to zero. This means that it is possible for contraband to be found but have no recorded amount to identify its type. Further, while for most observations we do know the general type of contraband found (ex. drugs or alcohol), we do not know the precise type of contraband found (ex. guns versus knives, or heroin versus marijuana).

Due to these complications, we standardize how we treat what contraband is found and the quantity found. To do so, we first calculate the percentile of the quantity found by category for each male driver in our subsample. Then we identify: (1) how many types of contraband are found based on the number of categories where a real quantity is listed; (2) then, for the category the driver ranks highest, the amount of contraband found. We then collapse the resulting percentiles into three equally sized bins. Additionally, for those drivers for whom no specific amount is recorded but contraband *is* found, a fourth bin is created. Further, the number of types of contraband found are collapsed into three categories: unidentified, one type found, and more than one type found.

Based on each of these variables and constructed categories, we collapse the data and calculate the proportion of male drivers who are arrested. This results in 11,182 "profile" observations of black drivers, 3,215 of Hispanic drivers, and 13,697 of white drivers. Figure 6 shows a stacked histogram of the proportion arrested in each profile, with bars stacked by race: Coral indicates black

[10] Safety stops include speeding violations, stop sign/light violations, driving under the influence, and safe movement violations. Investigatory stops include equipment violations, regulatory violations, not wearing a seat belt, criminal investigations, and other types of stops not listed. Checkpoint stops are excluded from the analysis.

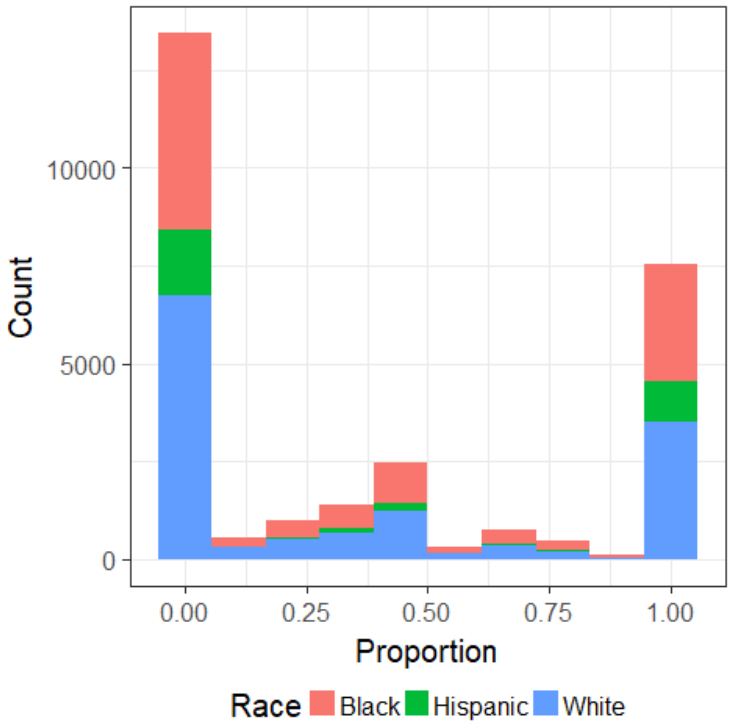

Figure 6: **Stacked Histogram of the Proportion of Male Drivers Arrested if Contraband is Found:** Observations are the proportion of male drivers arrested following a search and contraband being found for each agency, during a given time window, falling into a specific age group, where a specific amount of contraband was found, and following either an investigatory or safety stop. There are 11,182 observations of black drivers, 3,215 of Hispanic drivers, and 13,697 of white drivers.

drivers (the top color), green indicates Hispanic drivers, and blue (the bottom color) indicates white drivers. As can be seen, the distribution of proportion arrested is clearly bimodal, with profiles for which the proportion arrested was either very low or very high.

Additionally, the distribution of profile types appears to be very unbalanced across races. The top panel of Figure 7, which shows the observed differences in standardized means across all of our variable categories (i.e. the differences in standardized observed proportions of drivers with a given profile component), indicates that racial groups are generally discernibly different with respect to the contextual and demographic characteristics we have discussed. This poses a possible problem for the estimation of a regression predicting the proportion arrested as a function of race, as it becomes harder to attribute differences in arrest rates to race rather than to the things that covary with it in a way that is not heavily dependent on the model we specify. To address this problem, we pre-process the data by calculating weights designed to minimize these types of imbalances in

covariate distributions.

To obtain these weights, we estimate covariate balancing propensity scores (CBPS) that maximize balance in the distribution of covariates across race-ethnicity groups (Imai and Ratkovic, 2014; Fong et al., 2018).[11] Using these propensity scores, we compute the inverse probability of the race "treatment" to use as a weight in the regression fit. The bottom panel of Figure 7 shows that the balancing exercise is, overall, fairly successful: not only is the variance in differences larger before balancing the data than after, but also the average difference is reduced substantially as well. This should reduce bias and model dependence in our subsequent analysis.

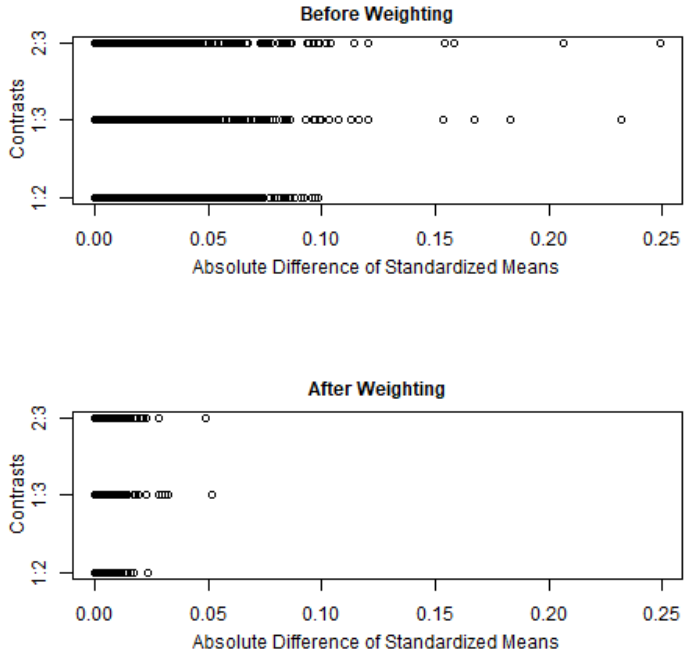

Figure 7: **Balance Plot**: Race is treated as the "treatment." Data is balanced with respect to agency, age, amount of contraband, stop purpose, number of contraband types, and the time of stop. In the figure, 1 indicates white drivers, 2 indicates black drivers, and 3 indicates Hispanic drivers.

Results

To test if race matters in the decision to arrest male drivers, we estimate a linear regression. The regression predicts the proportion arrested following a search where contraband is found. An

[11] We used the `CBPS` package in `R` to do so.

alternative interpretation of these regressions is that we are predicting the probability that an individual with a given profile is arrested.[12] As a reminder, each outcome is the proportion of male drivers arrested of a given race and age group found with an amount of contraband and a specific number of types of contraband at a specific time of day and following a given type of stop by a specific agency. The result is 28,094 observations. Additionally, observations are weighted by the inverse probability of "treatment" (viz. race) to balance the data set. Table 1 shows the results of the analysis, with standard errors calculated using heteroskedastic robust standard errors with the HC0 estimator.[13]

If the same trends are observed in the previous section on search thresholds are repeated here, we expect to see the coefficients associated with the black driver and Hispanic driver variables to be positive and statistically significant. This is precisely what we find. Black male drivers have an estimated probability of being arrested 4 percentage points higher than similarly situated male white drivers; Hispanic males face a 5 percent increased estimated probability of being arrested, in other wise similar circumstances. Each is statistically significant at the 0.05 level. By "similar circumstances," we mean cases that are similar with respect to the amount of contraband found, the age of the driver, and the time, purpose, and police agency associated with the stop.

What does this mean substantively? And how does the amount of contraband found influence arrest rates? To further illustrate what these results tell us and highlight how the amount of contraband found influences the probability of arrest, we turn to Figure 8, which presents the expected probability of being arrested following a search where contraband is found by race and amount found. When calculating the expected probabilities, the age category (26 to 35 years old), stop purpose (investigatory stop), number of types of contraband (one type), and time of the stop (between 5 pm and 10 pm) are held constant. Additionally, the stop is assumed to have occurred in Raleigh, NC by the city police department. In the figure, bars are grouped by amount of contraband found, while bars within the groups indicate driver race: coral bars (those on the left of the cluster) indicate black drivers, green bars (those in the middle) indicate Hispanic drivers, and blue bars

[12] Because there are values that are zero and are one in the data set, there may be some concern about whether OLS is the appropriate method to use, because the dependent variable is bound by zero and one. One way to test whether this poses an issue for the data at hand is to evaluate how many predicted values fall below zero or above one. In our case only 278 (out of 28,094) are either below 0 or above 1. In other words, just over 99% of the observations are predicted to fall in the identified range, suggesting that the use of what is effectively a linear probability model remains acceptable.

[13] To do so, we used the `sandwich` package in `R`.

Table 1: Estimating Proportion Arrested as a function of Race, Using Weighted Least Squares

	Model 1
Intercept	0.74*
	(0.07)
Black Drivers	0.04*
	(0.01)
Hispanic Drivers	0.05*
	(0.01)
Age: 26-35	0.07*
	(0.01)
Age: 36-45	0.10*
	(0.01)
Age: 46-55	0.10*
	(0.02)
Age: 56-65	0.06
	(0.04)
Age: 66-75	0.15*
	(0.07)
Age: 76-85	−0.05
	(0.10)
Age: 86-95	0.04
	(0.15)
Contra. Size: Middle	0.04*
	(0.01)
Contra. Size: Large	0.10*
	(0.01)
Contra. Size: Tiny	−0.18*
	(0.01)
Investigatory Stop	−0.03*
	(0.01)
1 Type of Contraband	−0.22*
	(0.01)
Hour: 5-10 am	−0.05*
	(0.01)
Hour: 11 am-4 pm	−0.01
	(0.01)
Hour: 5-10 pm	0.03*
	(0.01)
Agency Fixed Effect	Y
R^2	0.19
Adj. R^2	0.18
Num. obs.	28,094

Note: Regressions include proportion arrested following a search where contraband is found for male drivers. Covariate-balancing inverse probability of treatment estimates used as weights.

*$p < 0.05$

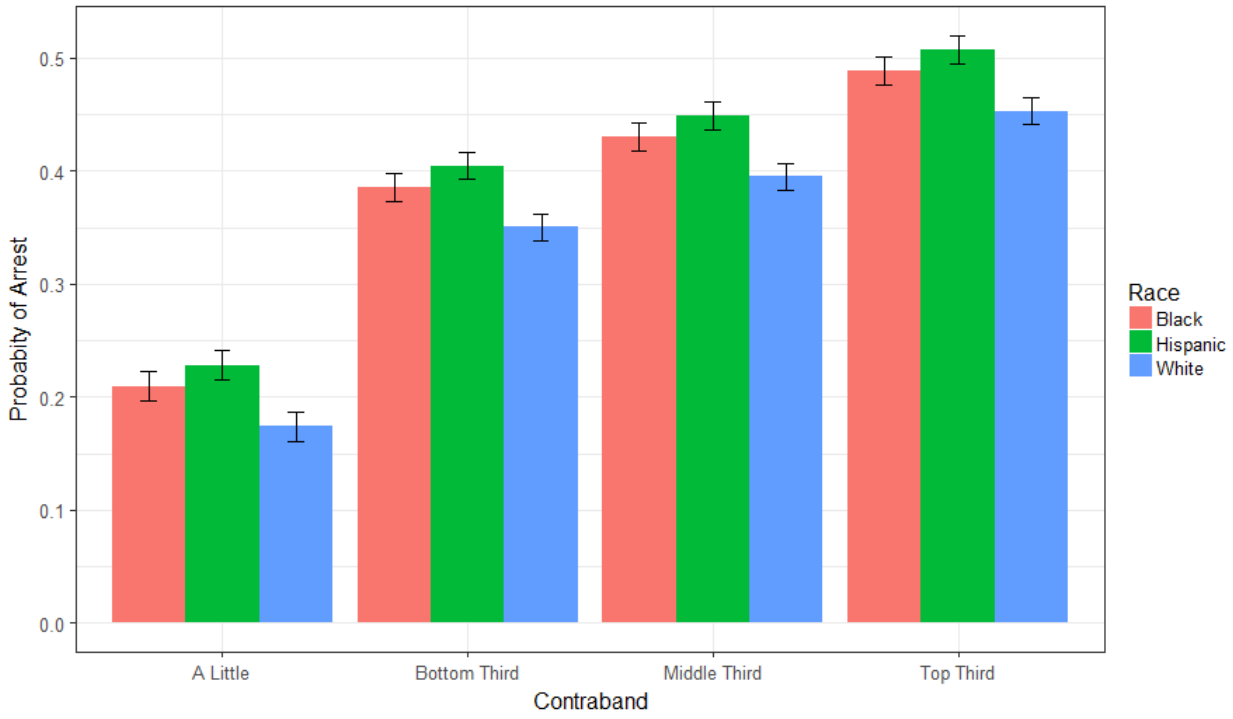

Figure 8: **Expected Probability of Arrest Following:** Age category held at 26-35, stop purpose is an investigatory stop, one type of contraband is found, and the stop occurred between 5 and 10 pm in Raleigh, NC. Expected probabilities based on the regression presented in table 1.

(those on the right) indicate white drivers.

In Figure 8, we can see that as the amount of contraband increases from a little or negligible amount[14] to the maximum category or those found with enough contraband to fall into the top third, the probability of arrest increases. The expected probability of arrest increases from under 25% to around 50% for both black and Hispanic drivers and from just over 20% to just over 45% for white drivers. Additionally, the most pronounced jump in the probability of being arrested occurs between a negligible amount found to an identifiable quantity of contraband.

The regression results from Table 1 also show that in addition to driver race and the amount of contraband found, several other factors also influence the chances of being arrested. These predictors include age, the purpose of the stop, the number of the types of contraband, and the time of day—even when the data has been balanced. At first, a few of the relationships emerging from this regression might appear counter intuitive. For instance, as the driver gets older (up

[14] As a reminder, these are amounts that appear as zeros across all possible categories in the data set due to such small quantities being recovered.

until the age of 75) the probability of being arrested following contraband being found increases.[15] Similarly, those stopped for investigatory rather than safety reasons are less likely to be arrested. And drivers are *less* likely to be arrested following a search where contraband is found between the hours of 5 and 10 am as compared to those stopped between 11 pm and 4 am, while stops occurring between 5 and 10 pm are *more* likely to result in arrest.

In isolation, any one of these findings might elicit surprise: after all, one might reasonably expect those who are found with contraband during the wee hours of the night, following an investigatory stop, and are younger rather than older drivers would be more likely to be arrested. However, when isolating the analysis to only those that have been searched and who were found to be holding contraband, these results become more plausible. If the stopping officer is in a situation with low discretion, an arrest—the expected outcome in all the cases we consider for this analysis—is more likely than in a situation where an officer has more discretion. For example, take the result for stop purpose: investigatory stops by definition afford officers a greater level of discretion than safety stops. If contraband is found following a search during a safety stop—when discretion is limited—arrest is more likely, whereas if its found following a search during an investigatory stop—when discretion is greater—arrest is less likely.

4 Discussion

We have explored racial differences in the odds of search and the odds of discovery of contraband following a routine traffic stop, based on millions of observations from North Carolina from 2002 through 2016, using a methodology that allows us to estimate the latent thresholds of suspicion that police officers use when making these stops. We showed the difference between these latent threshold models and a simple measure of the contraband hit rate, previously more common in the literature, but subject to inferential problems based on the issue of infra-marginality. Controlling then for these issues, we documented significant race- and gender-based disparities in the latent thresholds used, with some groups significantly under-policed and others over-policed. The rates at which drivers of different identity groups were subjected to fruitless search were particularly striking, with much higher false positive searches among minority male drivers. In sum, the accuracy of police

[15] The age reference category is those between 16 and 25.

perceptions of criminal suspicion differ dramatically and systematically by race and gender.

After having demonstrated these differences, we then turned to a matching exercise to ask a simple question: When two drivers are found in otherwise identical circumstances, for example having been stopped for speeding by an officer in a given police department, at a similar time of day, searched, and found with a similar amount of contraband, and sharing the same age and gender with another driver, but differing only by race, what are the differences in the odds of arrest? In this analysis, covering more than 26,000 profiles precisely matched in this way, black drivers appear to have a probability of arrest that is 4 percentage points higher, and Hispanic drivers have a probability of arrest that is 5 percentage points higher, than to their white counterparts.

Police face a difficult informational task. In a short encounter, they must reach an assessment of the odds of criminality. Our data show that falsely assuming criminality, like falsely assuming law-abidingness, differs systematically. Some groups are systematically under-policed, as evidenced by the data, while others—minority males in particular—are systematically subjected to more intrusive police action than is warranted. This includes the decision to search as well as the decision to arrest. Minority male drivers are searched with a lower apparent level of suspicion than whites, and arrested at higher rates, given otherwise identical circumstances.

References

Ayres, Ian. 2002. "Outcome tests of racial disparities in police practices." *Justice research and Policy* 4(1-2):131–142.

Baumgartner, Frank R, Derek A Epp, Kelsey Shoub and Bayard Love. 2017. "Targeting young men of color for search and arrest during traffic stops: evidence from North Carolina, 2002–2013." *Politics, Groups, and Identities* 5(1):107–131.

Baumgartner, Frank R., Epp Derek A. and Kelsey Shoub. 2018. *Suspect Citizens: What 20 million traffic stops tell us about policing and race.* Cambridge University Press.

Becker, Gary S. 2010. *The economics of discrimination.* University of Chicago press.

Brewer, Marilynn B and Layton N Lui. 1989. "The primacy of age and sex in the structure of person categories." *Social cognition* 7(3):262–274.

Devine, Patricia G. 1989. "Stereotypes and prejudice: Their automatic and controlled components." *Journal of personality and social psychology* 56(1):5.

Devine, Patricia G and Andrew J Elliot. 1995. "Are racial stereotypes really fading? The Princeton trilogy revisited." *Personality and social psychology bulletin* 21(11):1139–1150.

Epp, Charles R, Steven Maynard-Moody and Donald P Haider-Markel. 2014. *Pulled over: How police stops define race and citizenship.* University of Chicago Press.

Fiske, Susan T. 1993. "Social cognition and social perception." *Annual review of psychology* 44(1):155–194.

Fong, Christian, Chad Hazlett, Kosuke Imai et al. 2018. "Covariate balancing propensity score for a continuous treatment: application to the efficacy of political advertisements." *The Annals of Applied Statistics* 12(1):156–177.

Friedman, Jerome, Trevor Hastie and Robert Tibshirani. 2001. *The elements of statistical learning.* Vol. 1 Springer series in statistics New York, NY, USA:.

Imai, Kosuke and Marc Ratkovic. 2014. "Covariate balancing propensity score." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76(1):243–263.

Lerman, Amy E. and Velsa Weaver. 2014. *Arresting citizenship.* University of Chicago press.

Pierson, Emma, Sam Corbett-Davies and Sharad Goel. 2017. "Fast threshold tests for detecting discrimination." *arXiv preprint arXiv:1702.08536* .

Simoiu, Camelia, Sam Corbett-Davies, Sharad Goel et al. 2017. "The problem of infra-marginality in outcome tests for discrimination." *The Annals of Applied Statistics* 11(3):1193–1216.

Welch, Kelly. 2007. "Black criminal stereotypes and racial profiling." *Journal of Contemporary Criminal Justice* 23(3):276–288.