Search Distaste at the Highway Stop Level

Jonah B. Gelbach*
University of Pennsylvania Law School

EXTREMELY PRELIMINARY, WITH APOLOGIES FOR CITATION FAILURES

Current version:July 23, 2018
First version: [PENDING]

Abstract

A substantial literature tests for racial discrimination in highway searches of stopped drivers using data on search behavior, "hit rates", or both. This paper joins that literature by applying an insight of Arnold, Dobbie, & Yang (forthcoming, "ADY"). Writing in a different context—pre-trial release—ADY show that the standard Beckerian outcome-test framework can be empirically implemented using instrumental variables related to judge leniency in bail setting. I apply this idea to the highway search context, developing in detail a model of police behavior while taking driver behavior as given. The model implies that the search distaste of an officer in a stop equals the slope of the unconditional hit rate function—i.e., the hit rate among all stops, including those without a search—with respect to the probability of search. ADY also obtained this relationship, and they observed that it is connected to the marginal treatment effect and local instrumental variable parameters developed by Heckman & Vytlacil (1999) and Heckman & Vytlacil (2005).

I extend ADY's results and other work in the highway search literature in several ways. First, I derive three specification tests of the model. One of these, which does not require estimating search distaste values, is a modest generalization of a result in Anwar & Fang (2006), but I believe the other two are new. Second, I provide empirical estimates implementing the first specification test. I show that Anwar & Fang's (2006) data pass this test resoundingly, but that highway stops data from Harris County, Texas, do not. Third, I discuss conditions under which search distaste values are parametrically or nonparametrically identified at the highway stop level. Fourth, I discuss how to estimate these values (and how to use the estimates to measure racial discrimination). Fifth, I will (later) present estimates of search distastes.

*For helpful comments and suggestions, I thank David Abrams, Paul Heaton, and Jon Klick. I thank the Quattrone Center for the Fair Administration of Justice at the University of Pennsylvania Law School for grant funding.

1 Introduction

There is a substantial literature directed at testing whether police choose which stopped drivers to search in a racially discriminatory manner. Various authors have advocated and used tests involving police search behavior or the outcomes of searches to answer this question (see the discussion in section 2 for some examples). This paper shows that when information on both variables is available, we can learn considerably more about police officers' search tastes than seems to have been thought.

My approach is inspired by an application of Becker's (1957) outcome approach by Arnold, Dobbie & Yang (Forthcoming*b*, "ADY"). ADY consider a different substantive context from roadside searches—whether bail judges discriminate against black defendants. But the theoretical and econometric ideas they introduce map directly into the highway-search context. That said, their theoretical discussion takes place at a relatively high level, and there is some value to developing a more explicit analytical framework.[1]

I develop such a framework in this paper. The framework is simple enough that it takes drivers' behavior as given, focusing on a kind of reduced form in which only police officers' choices are made endogenously. Even this simple framework is general enough that it includes as special cases all the leading models of highway searches. Further, my framework yields a simple testable restriction, that the officer-level search rate and officer-level hit rate should covary negatively. Anwar & Fang (2006) also provide such a test. Although the two tests are closely related, mine is somewhat more general. Using the same Florida data that Anwar & Fang (2006) did, I test this restriction and find results that are in line with the model: officer-level search and hit rates do covary negatively, even after including a number of covariates.

A second contribution of the paper is that it clarifies the requirements for identification of police officers' search tastes at the driver-by-officer-by-stop level. I show directly that officers' tastes are identified under very weak conditions on the available data for any driver-officer-stop combination for which search occurs with positive probability, whenever (i) the probability of search is continuously distributed in the sample, or (ii) there is enough parametric structure to allow extrapolation across values of the search probability.

These conditions will resonate for readers who are familiar with recent development in the treatment effects literature. I show that the stop-and-search behavioral model I discuss can be represented as a generalized Roy model (see, e.g., Heckman & Vytlacil (2005) and Heckman & Vytlacil (1999)), also known as a potential outcomes model (see, e.g., Imbens & Angrist (1994)). This demonstration was also inspired by ADY, who also discuss the link between their framework and various papers that develop the econometrics of the generalized Roy model.[2]

[1] ADY also provide an alternative interpretation of their estimates, as inaccurate racial stereotypes rather than animus. I do not engage this alternative in the present paper.

[2] That said, ADY's method of estimating marginal treatment effects imposes considerable structure on the data generating process, as discuss in Appendix A. Further, as I discuss below, their results for black

I also discuss, in section 6, how to estimate officer search tastes under various assumptions about the structure of both officer search tastes and the driver misconduct-probability distribution. The nonparametric estimation discussion in this paper is a contribution relative to ADY, who impose strong functional form restrictions to justify their estimation approach. I also discuss how estimates of search distastes—officers' disutility of searching—may be used to construct discrimination measures at the driver-by-officer-by-stop-characteristics level. From these, aggregate measures of discrimination also may be constructed, e.g., including the average level of discrimination against a group, the fraction of stops in which a group's members face officers with discriminatory preferences, and so on. This is a big advance on the scope of information that can be gleaned from earlier tests for discrimination.

THIS VERSION OF THE PAPER DOES NOT YET INCLUDE ESTIMATES OF SEARCH TASTES.

A final contribution is to derive two additional restrictions the behavioral model makes on observed search distastes when driver and stop characteristics are held constant. First, the rate of observed driver misconduct among those drivers who are actually searched must be no greater than the average search distaste for officers in the stops that lead to searches. Second, unless hit probabilities are constant for the given driver and stop characteristics, officers' search distastes should be a decreasing function of the probability a search will occur. These model implications provide simple post-estimation specification tests.[3]

Section 2 kicks off the rest of this paper with a discussion of previous literature. Section 3 discusses the data I use. For Florida, this is the data set posted at the *American Economic Review* website by Anwar & Fang (2006). For Harris County, Texas, it comes from the Stanford Open Policing Project website.[4] Section 4 introduces the basic model of police search choices and derives its testable prediction for officer-level search and hit rates, and Section 5 presents empirical results for that test. The Florida data pass the test, but the Harris County data fail it. Section 6 then demonstrates that search distaste values are identified for stops whose driver, officer, and physical characteristics are associated with positive probability of search in equilibrium. Section 7 [WILL PRESENT] ~~presents~~ estimates of these search distaste values. Section 8 concludes.

2 Previous Literature

Early approaches to testing for discrimination focused on search rate differences by race, using statistical methods to adjust the racial difference for variation in covariates that might also affect search decisions; this approach is sometimes known as the benchmark approach.

defendants seem inconsistent with a basic fact about their data.

[3]As discussed below, ADY's estimates for black defendants in the pre-trial release context fail both tests. Their estimates for white defendants appear to pass the first test but fail the second over a broad range of search probabilities.

[4]See Pierson, Simoiu, Overgoor, Corbett-Davies, Ramachandran, Phillips & Goel (2017) for a description. The Project's data are described and made publicly available at https://openpolicing.stanford.edu/data/.

Because police officers might observe important characteristics that researchers don't, the benchmark approach raises omitted variables bias concerns: perhaps minority-group drivers are searched more often because they are more likely to be carrying contraband.[5] In that case greater search rates might be legally and substantively justifiable as being the result of targeting police resources effectively.[6]

A more recent literature investigates racial differences in the "hit rate"—the share of searches in which contraband is found, conditional on the occurrence of a search. If minority-group drivers are discriminated against by officers due to animus, officers will search them even when the probability of finding contraband is lower than for white drivers. This means minority-group drivers are being held to a higher standard than whites, which would show up in aggregate data as a lower hit rate for minority-group drivers.

The hit rate approach promises to leverage data on optimal choices to test for discrimination, obviating the problem of controlling for confounders. But the approach comes with a well known practical hurdle: inframarginality. Optimal choice usually involve variables at the margin, whereas easily observable data typically involve averages. When the two diverge in unobservable ways, the hit rate test breaks down.[7] To deal with this problem, Knowles, Persico & Todd (2001, "KPT") construct a two-sided game-theoretic model whose startling result is that there *is* no inframarginality problem. In their model, every driver of a type that is searched with positive probability will carry contraband with the same probability. But if there is taste-based discrimination against a given type of driver, then in equilibrium drivers of that type will carry contraband at a lower rate. So testing for discrimination against that type is as simple as testing the null hypothesis that the type's hit rate equals the hit rate for other types. The alternative hypothesis is that the search rate is lower for the type against whom discrimination allegedly operates. Outcome tests based on similar reasoning have been used in other settings, too, including Ayres & Waldfogel (1994); mortgage lending [CITES], and sentencing [CITES]. Ayres (2002) provides a general discussion of outcome tests (especially as applied to policing), including challenges related to inframarginality.

After KPT, a literature grew up suggesting various reasons why the inframarginality problem might occur after all. Both Anwar & Fang (2006) and Antonovics & Knight (2009) construct models in which officers' search distastes potentially vary across both officer race

[5]Simoiu, Corbett-Davies & Goel (2017, p. 1194–95) discuss a number of papers taking the benchmarking approach, as well as the "qualified pool"/"denominator" problem that arises from researchers' inability to observe all the determinants of decision makers' choices. Grogger & Ridgeway (2006) seek to avoid this problem via the "veil of darkness" by asking whether racial differences in the probability of being stopped differ across the level of daylight, on the theory that the decision to stop a driver can't depend on driver race if it is too dark for police to observe it; they conclude that data for Oakland, CA, do not reveal racial discrimination. Horrace & Rohlin (2016) point out that some parts of urban areas are well-lit at night and offer refined tests based on streetlight location, whose results for Syracuse, NY, indicate that there is discrimination. Kalinowski, Ross & Ross (2017) offer further discussion of the VOD test.

[6]It is a legal rather than economic or empirical question whether such statistical discrimination is lawful, and I do not engage that issue in this paper.

[7]See page 1194 of Simoiu et al. (2017) for a simple and clear example of inframarginality.

and driver race (KPT ruled this out by assumption).[8] KPT's constant hit rate result breaks down in both models. Persico & Todd (2006) subsequently offered a modified model that allows for some variation in officer tastes, and they showed that KPT's constant hit rate result emerges in that model's unique equilibrium.

There appear to be two reasons for the difference in the predictions of the KPT and Persico & Todd (2006) models, on the one hand, and those constructed by Anwar & Fang (2006), Antonovics & Knight (2009), as well as the one I offer below. One distinction is that Persico & Todd (2006) assume officers face a binding resource constraint; Antonovics & Knight (2009) argue that this assumption explains the difference in equilibrium predictions. Another distinction involves the process that matches drivers to officers. Persico & Todd (2006) write that Antonovics & Knight (2009, p. F363) "implicitly assume that drivers are randomly matched with a small set of police officers," creating "friction in the 'arbitraging process' that the police [department] follows," so "that not all drivers in a high-risk category need be searched."[9]

One reason to doubt the constant hit rate assumption is that driver behavior during a police stop might differ according to whether the driver is carrying contraband. As Anwar & Fang (2006, p. 133, footnote 20) observe, officers decide whether to search in part based on how stopped drivers behave in answering officers' roadside questions.[10] Drivers do not choose this behavior before they know whether they will be searched, so it does not fit well into the models of KPT or Persico & Todd (2006).[11]

The important point for my purposes is that the constant-hit rate prediction of the KPT and Persico & Todd (2006) models forms the basis for a simple test that can distinguish between those models and alternatives such as the one I offer. If hit rates are constant given observable characteristics, then officers who search a high share of stopped drivers should have the same hit rates as those who search a low share. If drivers' misconduct probability varies in ways that officers can observe but researchers cannot, then the general model below

[8] See also Sanga (2009), who reevaluates the evidence in KPT's original paper with some new data and empirical extensions. Other theoretical analyses appear in Dharmapala & Ross (2004), Bjerk (2007). Pierson et al. (2017) provide a hit rate analysis based on over 100 million police stops in 31 states, as well as results from a Bayesian approach to estimating the equivalent of search distaste at the race-by-police-department level; see also Simoiu et al. (2017), who provide a similar but more detailed analysis for North Carolina only. The latter paper has a particularly useful demonstration of how inframarginality can cause hit rate tests to fail.

[9] A questionable aspect of Persico & Todd's (2006, p. F356) model is that it appears to entail an equilibrium in which each police officer concentrates all his searches on exactly one type of driver.

[10] The instructions for officers on how to fill out the Florida Highway Patrol Traffic Stop Data Report (HSMV 62150) list "Physical/Visual Indicators" as one rationale for an officer's decision to request a consent search. In the Florida data used by Anwar & Fang (2006), this rationale is a frequent one listed as the reason for search (ADD DETAILS). Driver behavior is the example the instructions provide for such indicators: "For example, the subject no longer makes eye contact during the roadside interview or the subject crosses their arms when answering certain questions (becoming defensive)." Page 5.

[11] Unlike Knowles et al. (2001) and Persico & Todd (2006), I do not derive a two-sided equilibrium. In that sense my analysis should be viewed as reduced form in nature. I note that Anwar & Fang (2006) provide a two-sided equilibrium analysis of a model with a single observable type of driver and two officer types. They show the existence and uniqueness of equilibrium.

predicts that hit rates and search rates should be negatively associated. This is a slight generalization of the second sentence of Anwar & Fang's (2006) Proposition 3 (see page 136, as well as discussion at 138-139). Where I add value is in the use of officer-level measures of the search rate,[12] which allows me to use standard statistical testing of the constant hit rate prediction, whereas Anwar & Fang simply observe that certain statistics' ranks fit their predictions without using a formal test. I report empirical results for this test in section 5.

To my knowledge, the outcome test literature on highway searches has not used officer-level data. At most, the papers cited above have disaggregated data within cells defined by the driver's race and officer's race (Anwar & Fang (2006) and Antonovics & Knight (2009)) or by the driver's race and the officer's police department (Simoiu et al. (2017)). In this paper, I construct a simple behavioral model general enough to nest the models underlying KPT, Anwar & Fang (2006), Antonovics & Knight (2009), and Simoiu et al. (2017) and Pierson et al. (2017). The key objects of this model are the distribution of driver misconduct probability (discussed more below), and officer-level search distaste. Search distastes generally are identified for every type of stop—meaning every combination of driver, officer, and physical stop characteristics—that results in a positive probability of search in equilibrium.[13]

This is important. Anwar & Fang (2006) and Antonovics & Knight (2009) each note that when their results indicate discrimination is present, they are unable to identify which type of driver is the victim of it. The approach below allows one to answer that question.

The approach can be implemented empirically provided there is data on both (i) whether a stop results in a search and (ii) whether a stop results in a successful search. My theoretical analysis involves not the traditional (conditional-on-search) hit rate function for stops, but rather the unconditional hit rate function, although the two are of course related. The econometric results in this paper are driven by the structure of how the search function, the unconditional hit rate function, and officer search distaste interact.

As noted above, my approach and results analysis is inspired in important part by the insightful analysis in Arnold et al. (Forthcoming*b*, henceforth "ADY"). ADY study a different empirical setting: whether there is racial discrimination in bail setting by judges in Philadelphia and Miami. However, the behavioral model that underlies the part of their paper that focuses on taste-based discrimination is isomorphic to that presented here. The primary difference is in approach. ADY provide a relatively high-level discussion of the behavioral model, whereas I get into the details. ADY also link their econometric analysis to the marginal treatment effects ("MTE") literature (key citations are Heckman & Vytlacil (1999) and Heckman & Vytlacil (2005)). As their discussion implies, and as I explain in

[12]This is analogous to Arnold et al.'s (Forthcoming*b*) use of judge-level leniency measures in their study of pre-trial release and racial discrimination.

[13]This statement holds essentially without qualification when the propensity score has a continuous distribution. When it has a discrete distribution, identification is limited to points that have support in the propensity score distribution; see **?** for a discussion of identification with discrete instruments. ADD MORE HERE.

detail here, the behavioral model can be represented by the generalized Roy model that undergirds the MTE framework.[14] So identification results and estimation approaches from the MTE literature apply to the stop-and-search behavioral model.

An additional contribution in this paper is to point out some ways in which ADY's estimation method implicitly imposes considerable structure on officer-level search distaste parameters. Because ADY study a different topic—pre-trial release, rather than highway searches—I will not dwell extensively on their empirical results. I note, though, that for black defendants, both their two-stage-least-squares estimates and their MTE results imply that bail judges must be acting in a way that appears inconsistent with their data. Both results indicate judges place zero value on black defendants' pre-trial freedom. If that were the case, then judges would not release black defendants with positive probability of pre-trial rearrest. Yet black defendants in their data who are released before trial–who make up most of the sample of black defendants–are rearrested before trial roughly one-fourth of the time.[15]

One possible explanation for this pathological result would be that the behavioral model is a poor fit for the pre-trial release context. Bail judges do not decide simply whether to detain or release defendants before trial. Rather, they set a bail amount, which then helps determine whether a defendant will be released. But this mismatch may be unproblematic, e.g., if judges set bail to target release probability in such a way that the underlying behavioral structure still yields a data generating process that fits within the ADY's statistical framework. An alternative explanation involves ADY's estimation methods. As I discuss in Appendix A, these methods implicitly impose considerable structure on the related between officer's search distaste and the observable data. If ADY's structural assumptions are inappropriate, then that could explain their strange estimation results for black defendants.

One contribution of this paper is to point out alternative functional form assumptions under which search distaste parameters may be estimated using parametric or semi-parametric methods. My initial attempts to estimate the model under stuch restrictions led to estimates of the distribution of search distaste parameters that don't match observed hit rates. I thus conclude that if the behavioral model is acceptable, it likely will need to be estimated without imposing such structure. Thus another contribution of the paper is to suggest a non-parametric approach when such functional form assumptions are unattractive.

Non-parametric estimation with multiple covariates is data-hungry, requiring large sam-

[14]This is also sometimes known as the Rubin causal model or the potential outcomes model; Imbens & Angrist (1994) provide a classic discussion of important aspects of this framework. Vytlacil (2002) establishes the equivalence of the generalized Roy model and the model underlying Imbens & Angrist (1994).

[15]Compare ADY's Table 4 2SLS estimate of the average judge-release taste parameter for blacks, which is almost exactly 0, to their Table 1 summary statistic showing that roughly a quarter of black defendants who are released pre-trial are arrested again before the trial. With respect to MTE results, the average reported in Table 4 is nearly identical to that for 2SLS; further, the graph of the MTE for black defendants in Figure II hovers around 0 over the entire propensity score range from 0 to 1. As for ADY's estimates for whites, their estimates yield an average MTE roughly equal to the rearrest rate among released whites. This also seems problematic, because the behavioral model predicts pre-trial release only for those with a rearrest probability *no greater than* than the social value the judge places on releasing the defendant. Thus one would expect to see an average MTE that is lower than the rearrest rate among released whites.

ple sizes to make up for the lack of parametric structure. Fortunately, there are large datasets on roadside searches. The first data source I use is the one used by Anwar & Fang (2006), which I downloaded from the American Economic Review's website. The data come from nearly a million traffic stops in Florida between January and November of 2001. The second data source comes from the Texas data posted by the Stanford Open Policing Project and used in Pierson et al. (2017).[16] I use data from several hundred thousand stops in Harris County, home of the city of Houston.

[NOTE: I HAVE NOT YET GENERATED THE NONPARAMETRIC ESTIMATES FOR THIS PAPER. THIS VERSION OF THE PAPER IS LIMITED TO SOME PRELIMINARY EMPIRICAL FACTS FOR FLORIDA AND THE ANALYTICAL DISCUSSION BELOW.]

3 Data

3.1 Florida data

For Florida, I use the data used by Anwar & Fang (2006).[17] Table 1 reports some basic summary statistics for this sample. The full sample has data on 906,339 stops. Of these, about 600,000 involved white drivers, with the remaining 300,000 or so split roughly evenly among black and Hispanic drivers.

The overall search rate was 1%. For both blacks and Hispanics, it was 1.3%, and for whites it was 0.8%. This difference is small in levels, but these figures imply that given a stop, blacks and Hispanics are searched about 60% more often than are whites.

The conditional hit rate varies considerably across race. It is greatest among whites, at 0.251. The rate for blacks is 0.209, and for Hispanics it is a much lower 0.115. Given the sample sizes, these figures are statistically significantly different for all pairwise comparisons. KPT's test would thus indicate discrimination in search decisions made against blacks and Hispanics relative to whites, and against Hispanics relative to blacks.[18]

About 70% of drivers were male, although this varies between 67% for blacks to 77% for Hispanics. Roughly 90% of officers conducting stops were male; this figure did not vary appreciably by driver race. Overall, 16% of officers were black, about 11% were Hispanic, and the other 73% were white. These statistics vary considerably across driver race, with officers more likely to be involved in a stop of drivers with the same race. For example, white officers conducted only about 57% of stops of Hispanic drivers; although that constitutes a majority of stops of Hispanics, it is 20 percentage points lower than white officers' share of stops of white drivers. Most stops—about 70%—occurred between 6 a.m. and 6 p.m.

[16] The Project's data are described and made publicly available at https://openpolicing.stanford.edu/data/.

[17] See files contained in https://www.aeaweb.ORG/aer/data/mar06_data_20040597.zip.

[18] Because search rates are so low, the unconditional hit rate is very low—0.002 for Hispanics and whites and 0.003 for blacks.

Roughly 10% of stopped vehicles were registered outside Florida; this figure was about half as great for vehicles with Hispanic drivers. Between 39% and 45% of stopped vehicles had at least one passenger.

3.2 Harris County, Texas, data

For Harris County, I use Texas stops data posted by the Stanford Open Policing Project and used in Pierson et al. (2017).[19] Table 4 reports summary statistics for this data set. The full sample has data on 618,497 stops that took place between 2006 and 2015 (the number of stops varies between a low of 48,000 in 2015 to a high of just over 75,000 in 2006). Of these, about a quarter each involved black or Hispanic drivers, and about 44% involved white drivers; the rest involved drivers of other racial backgrounds.

The overall search rate was 1.6%, varying between 1.4% and 1.5% for whites and Hispanics to 2.2% for blacks. Unlike the Florida data, the conditional hit rate in Harris County varies relatively little across race. It is nearly identical for blacks and whites (0.349 and 0.340, respectively), and just a few points lower (0.303) for Hispanics. The unconditional hit rate is 0.005 for Hispanics and whites and 0.008 for blacks. Roughly two-thirds of drivers are male, although there is some variation in this statistic across race. A bit more than half of stops occur between 6 a.m. and 6 p.m.

Unfortunately, less information is available for Harris County stops. I do not have data regarding trooper race, trooper gender, state of registration, or number of passengers (all of which are availabe in the Florida data).

4 Theory

In this section, I construct the basic model. It is general enough to nest those in KPT, Anwar & Fang (2006), Antonovics & Knight (2009), and Simoiu et al. (2017) and Pierson et al. (2017). The analysis yields the same basic methodological results as ADY, in addition to a testable prediction that may be used to distinguish my general approach (and that of Anwar & Fang (2006) and Antonovics & Knight (2009)) from KPT and Persico & Todd (2006).

Section 4.1 introduces the model. Section 4.2 defines the search rate and both the conditional and unconditional hit rate functions.

4.1 The basics

Consider a single officer who stops drivers according to a decision rule which is unmodeled. I will distinguish three types of driver characteristics: r denotes race, X denotes a vector of

[19] The Project's data are described and made publicly available at https://openpolicing.stanford.edu/data/.

characteristics that can be observed both by an officer who stops a driver and researchers, and m denotes a characteristic that is observable to officers but unobservable to researchers.

On stopping a driver, the officer observes information from which he forms a belief, $m \in [0, 1]$, about the probability that the driver is engaged in misconduct by carrying contraband. Think of this variable as a transformed version of the noisy signal that underlies Anwar & Fang's (2006) theoretical analysis.[20]

I assume that if the officer searches a driver, there is probability 1 that he will find any contraband that is present. When officer j stops a race-r driver with characteristics $X = x$, his disutility of searching the driver—his search distaste—is $\delta_{r,j}(x) \in [0, 1]$. Following the literature, I assume search distaste is measured in units of the officer's expected social cost of allowing a driver to carry contraband rather than intervening (whether to confiscate it, arrest the driver, or both), so this the benefit of finding contraband in a search is normalized to 1.

For officer j facing drivers with characteristics $X = x$, define animus against race-r_1 drivers, relative to race-r_2 drivers, as the x-specific difference in search distaste:

$$\Delta_j(x) \equiv \delta_{r_2,j}(x) - \delta_{r_1,j}(x). \tag{1}$$

This difference may in principle take any sign for any given officer j.

I assume the officer will search a driver whenever the probability of finding contraband exceeds the officer's disutility of searching, $m \geq \delta_{r,j}(x)$, and not otherwise.[21] If officer j harbors animus against race r_1 relative to r_2, he will search some r_1 drivers whom he would not search if they were of race r_2. The parameter $\Delta_j(x)$ measures just how great the scope for such disparate treatment is for type-x drivers stopped by officer j.

The average level of animus for officer j is $\overline{\Delta}_j \equiv E_X[\Delta_j(X)]$, with the expectation taken over the distribution of X for stops conducted by officer j. The average level of animus for type-x drivers is $\overline{\Delta}(x) \equiv E_j[\Delta_j(X)]$, where this expectation is taken over officers while holding non-race characteristics at $X = x$. The average level of animus over all drivers and officers is the expectation taken over both distributions: $\overline{\Delta} \equiv E_{X,j}[\Delta_j(X)]$. If animus measures are identified at the officer and/or x level, then of course one can also measure more than just averages; quantiles are possible, as well as the share of officers who harbor animus against a particular race, and so on.

Let $F_{M|X,r}$ be the *cdf* of misconduct probability across the distribution of stopped drivers given characteristics X and r. Allowing F_M to be non-degenerate is a critical feature that distinguishes approaches in the literature.

[20] In Anwar & Fang's framework, the conditional probability that a driver is carrying contraband is a strictly increasing function of the noisy signal they call θ. I therefore change nothing of substance when I work directly with the probability m rather than the signal that determines its value.

[21] Nothing would change if I assumed the officer randomizes or doesn't search when indifferent.

4.2 Hit rates and the search rate

For this section I will suppress the notation related to X and race. The probability that a search will turn up contraband among those drivers searched by a officer with search distaste δ is $E[M|M > \delta]$. Assuming that the misconduct probability distribution F_M is continuous, the probability an officer will search a randomly drawn driver among those he has stopped is[22]

$$\begin{aligned} s(\delta) &\equiv P(M > \delta) \\ &= 1 - F_M(\delta), \end{aligned} \tag{2}$$

and the hit rate for a officer with search distaste δ is

$$h(\delta) = \int_{\delta}^{1} m dF_{M|M>\delta}(m), \tag{3}$$

where $F_{M|M>\delta}$ is the conditional *cdf* of the misconduct probability among drivers searched by a officer with search distaste δ. By assumption, this distribution is continuous, so $dF_M(m) = f_M(m)dm$. Thus the conditional density among drivers searched by a officer with search distaste δ is given by

$$\begin{aligned} f_{M|M>\delta}(m) &= \frac{f_M(m)}{1 - F_M(\delta)} \\ &= \frac{f_M(m)}{s(\delta)}, \end{aligned} \tag{4}$$

Now define the unconditional hit rate as

$$\upsilon(\delta) \equiv \int_{\delta}^{1} m f_M(m) dm. \tag{5}$$

This is the total mass of successful searches for an officer with search distaste δ. Multiplying and dividing (5) by $s(\delta)$ shows the intuitive result that the unconditional hit rate is the product of the search rate and the hit rate given that there has been a search, to which I will sometimes refer as the conditional hit rate. Thus

$$\upsilon(\delta) = \int_{\delta}^{1} m f_M(m) dm = s(\delta)h(\delta). \tag{6}$$

Now suppose we have two officers, i and j, who have search distastes $\delta_i < \delta_j$. Officer i will search more drivers, so we have $s(\delta_i) \equiv s_i > s_j \equiv s(\delta_j)$. Because i is willing to

[22]Embedded in this treatment of officer behavior is the assumption that the officer does not face a time constraint preventing him from searching as many stopped drivers as he likes. This assumption follows the approach in Knowles et al. (2001), Bjerk (2007), and Antonovics & Knight (2009), but differs from that in Persico & Todd (2006). For more, see the discussion in Antonovics & Knight (2009) (at page 165, in footnote 5, and at page 167).

search those with lower misconduct probability, i's conditional hit rate will be lower than j's: $h(\delta_i) \equiv h_i < h_j \equiv h(\delta_i)$. Because officer i searches all the drivers officer j does, as well as some additional ones, i will have more successful searches in total as well. In other words, $u_i \equiv \upsilon(\delta_i) > \upsilon(\delta_j) \equiv u_j$.

We can collect the results so far in the following proposition.

> **Proposition 1.** If $F_{M|X}$ is nondegenerate, the search probability and unconditional hit probability, s and u, are decreasing functions of the search distaste δ. The hit probability, h, is an increasing function of τ.

An immediate empirical implication of Proposition 1 is that officers' search and conditional hit rates against drivers with a given (r, X) value will covary negatively. If instead the misconduct probability is constant given (r, X), as in Persico & Todd's (2006) equilibrium, then the conditional hit rate will not vary with officers' search rates. Thus vis-a-vis KPT's model, Proposition 1 makes a testable prediction about the relationship stop-and-search behavioral model developed above. Anwar & Fang (2006, p. XXX) derive what is essentially the same result, although Proposition 1 is stated in a more general way in that it does not depend on the race or ethnicity of officers. [ADD MORE DISCUSSION OF AF2006 ON THIS POINT, AS WELL AS ANTONOVICS & KNIGHT.]

There is some evidence available on this question from results reported in Anwar & Fang (2006). They show [XXX ADD]. In the next section, I offer my own estimates on the empirical relationship between the search and conditional hit rate functions.

5 The Empirical Relationship Between s and h

To test the sign of the relationship between the search and conditional hit rate functions, I generated OLS estimates of the relationship between a dummy for whether contraband was found in a search, among those stops in which a search was conducted, and the officer-level search rate among all other drivers of the same race as the driver in question who is stopped by the officer in question.[23]

Figure 1 plots estimated densities of the number of stops made by the officer in each stop, by driver race. To construct the number-of-stops variable for a given stop of a race-r driver by officer j, I count up the total number of drivers of race r who were stopped by officer j. I then use a kernel estimator to estimate the density of the number-of-stops variable, treating stops as the unit of analysis (thus officers with more stops automatically get more weight). The figure shows that there is considerable variation in the number of stops, and there are

[23] That is, I construct the officer-level search rate using a leave-one-out approach, to avoid any mechanical relationship. Let $D_{i(j,r)}$ be a dummy variable for whether race-r driver i is searched by officer j. Then the leave-one-out officer-level search rate for driver i' is $(N_{jr} - 1)^{-1} \sum_{i(j,r) \neq i'}^{N_{jr}} D_{i(j,r)}$, where N_{jr} is the total number of race-r drivers stopped by officer j. Notice this equals $(N_{jr} - 1)^{-1}(N_{jr}\overline{D}_{jr} - D_{i'})$, so the leave-one-out search rate can be calculated using only two runs through the data, rather than N.

a sizable number of stops made by officers who made large numbers of stops (this latter statement is somewhat less true for stops of black drivers).

Figure 2 plots kernel estimates of the density of the leave-one-out officer-level search rate variable among those stops in which there was a search. Most of the mass of this variable is concentrated below a search rate of 10% or so, but there is still a fair amount of variation even above that level.

5.1 Florida Estimates

Each specification includes a dummy for whether the driver was male, a dummy for whether the stop occurred between 6 a.m. and 6 p.m., a dummy for whether the car was registered outside Florida, a dummy for whether there was at least one passenger in the car, and a set of location dummies.

Table 2 reports the first set of these results, using county dummies for location. It shows a strong negative relationship between the probability of finding contraband in a search and the officer-level search rate. For black drivers who were searched, a one percentage-point increase in the officer-level search rate is associated with a 0.14 percentage-point drop in the hit rate. The association for Hispanics is about one-third greater than this, and for whites it is 40% greater. Among those drivers who were searched, and thus included in the OLS estimation, the standard deviation of the officer-level search rate variable is 18 percentage points (i.e., 0.18) for black drivers, 24 percentage points for Hispanic drivers, and 11 percentage points for white drivers. The estimates in Table 2 thus imply that a one standard deviation increase in the officer-level search rate is associated with a drop in the hit rate of 2.52 percentage points for black drivers, 4.44 points for Hispanic drivers, and 2.16 percentage points for white drivers. When compared to the overall hit rates (20.9% for blacks, 11.5% for Hispanics, and 25.1% for whites), the point estimates are somewhere between noticeable and quite substantial.

Under my model's alternative to the KPT null hypothesis of no variation in the conditional hit rate, the conditional hit rate varies negatively with the officer-level search rate. So one-sided critical values are appropriate for testing the KPT null against my model's prediction. The one-sided p-values for the estimates are below 0.03 for all three races.

Table 3 reports analogous estimates with district dummies used in place of county dummies. Compared to the specification with district dummies, these estimates are larger for blacks, roughly the same for Hispanics, and much greater for whites. In all three cases the coefficients are estimated with great precision.

As for the coefficients on the covariates in Table 2 and Table 3, driver's gender has only a small association with the hit probability for blacks and Hispanics. For whites, the estimates indicate that for male drivers a hit is 2.6 and 3.9 percentage points more likely; these are nontrivial figures.[24] The daytime-search coefficients are small or imprecisely estimated for

[24]For these coefficients, the appropriate hypothesis test is two-sided, because the stop-and-search behavioral model makes no prediction about their sign. The two-sided p-values for the two estimates against a

black and Hispanic drivers but are larger and marginally statistically significant for whites.[25] Out-of-state registration is associated with sizable reductions in the hit rate for all three driver races; this association is precisely estimated for Hispanics and whites, though not for black drivers. Finally, having at least one passenger in the vehicle is associated with substantial increases in the hit probability of a search; all the coefficients for this variable are precisely estimated.

Taken as a whole, these results provide strong support for the view that the hit rate and search rate are negatively related. The results are statistically significant against a null hypothesis of zero relationship, and they are substantial in magnitude. They hold up for all three races across a variety of specifications. Further, the results indicate that a number of other characteristics of officers, drivers, and stops also are related to hit probabilities.

These findings provide strong support for the idea that $F_{M|X}$ varies with X and, for at least some X characteristics, is a non-degenerate function. Thus, the estimates in Table 2 and Table 3 support the application of my stop-and-search behavioral model to the Florida data relative to the KPT and Persico & Todd (2006) models, which imply that the hit rate is constant for fixed (X, r) values.

5.2 Harris County Estimates

Table 5 reports OLS estimates of the hit rate/officer-level search rate for Harris County. For stops involving black drivers, the estimate is negative but estimated with terrible precision. For stops involving Hispanic and white drivers, the coefficient estimates are large in magnitude and positive. Both of these estimates are wrong-signed relative to the theory, and each is statistically significant at conventional levels.

This finding suggests the model is a poor fit for the Harris County data. This could be because police behavior in Harris County differs in important ways from the model's assumptions. It could also arise because I have access to so few X covariates for Harris County. If such variables are associated with both the distribution of driver misconduct probability, $F_{M|X}$, and police officers' search distastes, then it would be possible to observe the reults in Table 5 even if the model is an accurate one.

I note that the results in Table 5 also reject the KPT/Persico & Todd (2006) prediction that within race, hit rates will be constant for all those who are searched. One reason to conclude this is the positive and statistically significant value of the coefficient on the officer-level search rate variable for whites and Hispanics. A second reason is that several of the coefficients for driver gender and daytime stops are statistically significantly different from zero at conventional levels. This would not occur under the KPT/Persico & Todd (2006)

null of zero are 0.115 and 0.025.

[25]The two-sided p-values for the coefficients were 0.087 and 0.099. Recall that the sample size for white drivers is several times the sample sizes for black and Hispanic drivers. Thus the greater precision of the coefficient estimates for whites is to be expected, although of course the coefficient sizes may truly differ as well.

prediction.

Because the empirical results for Harris County fail the specification test discussed in this section, I shall not consider Harris County data further in this paper.

6 Identifying search tastes using the unconditional hit rate function

In section 6.1 I extend the basic model above to analyze the relationship between search distaste, the search rate, and the unconditional hit rate function. Then in section 6.3 I discuss some parametric assumptions under which search distastes are identified using information on the search rate function alone, as well as when information on the unconditional hit rate function are needed as well. In section 6.4 I discuss nonparametric identification of search distastes.

6.1 Search tastes and the slope of the unconditional hit rate function

Equation (6) may be used to relate the unconditional hit rate and the search rate to the distribution of misconduct probability via the equation $\upsilon(\delta) = \int_{\delta}^{1} m f_M(m) dm$. Differentiating with respect to δ yields

$$\frac{d\upsilon}{d\delta}(\delta) = -\delta f_M(\delta). \tag{7}$$

Because $ds/d\delta = -f_M(\delta)$, it follows that $d\upsilon/d\delta = \delta ds/d\delta$. The distribution of the misconduct probability F_M is strictly increasing, we may write a generic officer's search taste as a function of the officer's search rate:

$$\delta = t(s) \equiv F_M^{-1}(1-s). \tag{8}$$

Therefore we may also write a generic officer's unconditional hit rate as an indirect function of the officer's search probability, rather than his search taste, i.e., we may work with $\upsilon(t(s))$. By the chain rule, $d\upsilon/ds = (d\upsilon/d\delta)(dt/ds)$, so

$$\frac{d\upsilon}{ds}(t(s)) = \delta. \tag{9}$$

In words, for any given search rate s, the rate of change of the unconditional hit rate, as we move up the driver misconduct-probability distribution in a way that increases the search rate infinitesimally, exactly equals the search distaste of the officer who searches at

the given search rate. Further, because $t(s) = F_M^{-1}(s)$, this rate of change also equals the $(1-s)$-quantile of the distribution of misconduct probabilities. Thus the distribution of officer preferences is identified, over all values of s for which a stop has probability s of having a search, by knowledge of the joint distribution of search probability and either the conditional or the unconditional hit probability. An implication is that if it is possible to consistently estimate the probability of search and either the conditional or the unconditional hit rate, then it will be possible to estimate the distribution of officer preferences.

In addition, because the officer preference distribution maps one-to-one onto the distribution of driver behavior for those drivers who are ever searched, key features of the driver misconduct distribution are also identified.

6.2 Two more testable restrictions

The behavioral model says that an officer will search any driver with a misconduct probability above the officer's search distaste. This yields the following proposition.

> **Proposition 2.**
> The average probability of misconduct for actually searched drivers is at least as great as the average search distaste among officers in those stops that end in a search.

This means inframarginality may be used as the analytical basis for a specification test. For any set of driver-officer-stop combinations, the observed hit rate must be no greater than the average estimated MTE.

A final analytical result involves how officers' search distaste moves with the probability of search, holding $X = x$. Recall that the search rate function for an officer with search distaste δ is $s(\delta) = \int_\delta^1 f_{M|X}(m|X = x)$. Recall also that because this sets up a monotonic relationship between search probability and search distaste, we wrote $\delta = t(s)$ for the inverse search probability function. This yields a final proposition.

> **Proposition 3.**
> Holding $X = x$, the search distaste function has slope $-[f_{M|X}(\delta|X = x)]^{-1} < 0$. That is, in equilibrium, officers' search distaste is a decreasing function of search probability.

Proposition 3 also entails an empirical prediction: if the stop-and-search behavioral model above is correct, then a plot of the MTE for any fixed characteristics $X = x$ should be decreasing in the search rate for drivers and stops with $X = x$.

6.3 Two parametric examples

I now use the results above to show how officers' search tastes are identified in the equilibrium of the stop-and-search behavioral model. The two key that allow identification of officer

search tastes are (2) and (9). Writing $S(j,r,x)$ for the probability that officer j will search a race-r driver with characteristics $X = x$, I rewrite these equations as

$$\begin{aligned} S(j,r,x) &= 1 - F_{M|X,r}(\delta_{r,j}(x)|r, X = x) \\ \frac{d\upsilon}{ds}(t(s(j,r,x))) &= \delta_{r,j}(x). \end{aligned} \tag{10}$$

The first equation in (10) relates the conditional probability of search to observable features (j,r,x); it may be thought of as a propensity score equation. Suppose the functional form of $\delta_{r,j}(x)$ is known up to a finite set of parameters. Then identification of $F_{M|X,r}$ from information on the distribution of (r,j,X) and the realized search outcome is a familiar problem from the semi-parametric literature. If in addition $F_{M|X,r}$ is known up to a finite number of parameters, the problem becomes a conventional one of parametric identification. In both cases—i.e., whenever the functional form of $\delta_{r,j}(x)$ is known—it would seem that standard conditions might be enough for identification of $\delta_{r,j}(x)$ using only the propensity score equation.

I now consider two parametric examples. In the first, the propensity score provides enough structure to identify $\delta_{r,j}(x)$. In the second example, the propensity score does not provide enough information to identify $\delta_{r,j}(x)$, but this parameter nonetheless is identified by the slope of the unconditional hit rate function.

For clarity, I suppress notation related to race. I let $\tilde{Z}$ be an instrumental variable that shifts officer preferences, and I write $\delta_j(x) = \delta(x, \tilde{z}_j)$ for an officer with $\tilde{Z} = \tilde{z}_j$ during a stop with $X = x$. Write $Z = (X, \tilde{Z})$, and let $z_j = (x, \tilde{z}_j)$ be a particluar realization. I now consider two examples, each having

$$\begin{aligned} F_{M|X}(\delta|X = x) &= F_0(F_0^{-1}(\delta) - x\alpha_0) \\ \delta(x,\tilde{z}) &= g(x\alpha_1 + \tilde{z}\gamma). \end{aligned} \tag{11}$$

Notice that an observation with $x\alpha_0 = 0$ has $F_{M|X}(\delta|x\alpha_0 = 0) = \delta$. This is a normalization rather than a restriction: by Brouwer's fixed point theorem, $F_{M|X}$ must have a fixed point on $[0,1]$, and as long as there is a constant in x, we can choose its coefficient in α_0 so that the fixed point occurs where $x\alpha_0 = 0$. The content of the first equation in (11) is instead that x acts only to shift the location of $F_{M|X}$ relative to the δ-quantile of the F_0 distribution. Examples 1 and 2 just below differ only in their choices of "link" functions, F_0 and g.

Example 1 (Logistic-Normal Links)**.**
Suppose that F_0 is the logistic *cdf*, Λ, with inverse $\text{logit}(\delta) = \ln(\delta/(1-\delta))$, and suppose $g = \Phi$, i.e., the standard normal *cdf*.

$$\begin{aligned} F_{M|X}(\delta|X = x) &= F_0\left(\text{logit}(\delta) - x\alpha_0\right) \\ \delta(x,\tilde{z}_j) &= \Phi(x\alpha_1 + \tilde{z}_j\gamma). \end{aligned} \tag{12}$$

Let D be a dummy variable that equals 1 when there is a search and 0 otherwise. Sub-

stituting the second equation in (12) into the first yields

$$\begin{aligned} P(D=1|X=x,\tilde{Z}=\tilde{z}) &= 1-F_{M|X}(\delta|X=x) \\ &= 1-\Lambda\left(\ln\left(\frac{\Phi(x\alpha_1+\tilde{z}_j\gamma)}{1-\Phi(x\alpha_1+\tilde{z}_j\gamma)}\right)-x\alpha_0\right). \end{aligned} \tag{13}$$

The parameters α_1 and γ are identified except in pathological cases, and they can be estimated using conventional parametric methods. The search distaste for each stop could then be estimated directly using the estimates of α_1 and γ. Notice that these claims would generally hold even if $\gamma = 0$, i.e., there is in effect no officer-level instrumental variable. In such a case, identification comes from the nonlinear way that x enters (13).

Now consider a second example in which the nonlinearity avenue for identification is closed.

Example 2 (Logistic-Logit Links)**.** Assume that $g = \Lambda$, i.e., the δ link is logistic rather than normal. Then from (12) we have

$$\begin{aligned} P(D=1|X=x,\tilde{Z}=\tilde{z}) &= 1-F_{M|X}(\delta|X=x) \\ &= 1-\Lambda\left(x(\alpha_1-\alpha_0)+\tilde{z}_j\gamma\right). \end{aligned} \tag{14}$$

From (14) it is clear that the selection equation identifies $(\alpha_1 - \alpha_0)$ and γ. But it doesn't by itself allow us to disentangle x's role in officer j's preferences (α_1) from its role in the conditional misconduct distribution (α_0).

But all features of the Example 2 model still may be identified. The selection equation in (12) can be expressed as $\delta(x, \tilde{z}_j) = \text{logit}(1 - S(j, r, x)) + x\alpha_0$, of which only α_0 is not identified by the selection equation. Inserting this expression into the second equation in (10) and setting $s = S(j, r, x)$, we have

$$\frac{d\upsilon}{ds} = x\alpha_0 + F_0^{-1}(1-s). \tag{15}$$

Integrating both sides with respect to the search probability yields

$$\begin{aligned} P\left(Y=1|X=x\,,S(j,X)=s\,\right) = \\ x\alpha_2+(x\alpha_0)s+\int \text{logit}(1-s)ds, \end{aligned} \tag{16}$$

where Y is an indicator variable that equals 1 if a stopped driver is both searched and found to be carrying contraband, and equals 0 if the driver is either not searched or not carrying contraband.

The left hand side of (16) is the conditional expectation of the unconditional hit rate, given observables $(j, x, \tilde{z}_j)$. Because s is identified from the selection equation and the integral of the logit function has a known (and closed) form, α_0 and α_2 are identified as under standard conditions for the identification of linear regression parameters.[26] Thus both

[26]Namely, the random vector $(X, X \times s)$ must have an invertible covariance matrix.

the components of $\delta(x, \tilde{z}_j)$ are identified, so $\delta(\cdot)$ is as well.

This double-logit example above is only one instance of a more general class for which the identification analysis would wind up at (16). Whenever (11) holds and $g = F_0$, that will be the case.[27]

These two examples show that identification of officer j's search distaste in a stop of a race-r driver with characteristics x is generally possible when $F_{M|X}$ and $\delta(\cdot)$ have certain kinds of structure and are known up to parametric form.

6.4 Nonparametric identification of $\delta_j(x, \tilde{z}_j)$

It is also possible to identify officer j's search distaste without making particular structural assumptions on $F_{M|X}$ and $\delta(\cdot)$. Begin with the propensity score equation. For each stop, data are available on whether a stopped driver is searched, as well as on (j, x). Let D be an indicator variable that equals 1 if a stop results in a search and 0 otherwise. Because the search rate given (j, x) is $E[D = 1|j, X = x]$, $S(j, x)$ is identified whenever this conditional expectation is. Conditions for nonparametric identification of a function such as $E[D|j, X = x]$ have been widely discussed; CITES. When such conditions hold, we may treat $s = S(j, x)$ as known for purposes of identification analysis.

We know from the propensity score equation that $\delta(j, x) = F^{-1}_{M|X}(1 - s|X = x)$. Thus we have $\upsilon(t(s(j, x))) = \int F^{-1}_{M|X}(1 - s|X = x)ds$. Because $F_{M|X}$ is unknown, this integral cannot be calculated analytically. But notice that $\upsilon(t(s(j, x)))$ is the probability that a stop (i) of a race-r driver with characteristics x, (ii) by an officer with search probability $S(j, x) = s$, (iii) will result in a successful search. Letting Y be an indicator function equal to 1 if a stop ends in a successful search for contraband and 0 in all other cases (including no search), $\upsilon(s; r, x) = P[Y = 1|X = x, S(j, x) = s]$. Accordingly, $\delta(x, \tilde{z}_j)$ is identified whenever the derivative of $P[Y = 1|X = x, S(j, x) = s]$ with respect to s is identified.

Integrating this derivation with respect to s and using the relationship between $\delta(\cdot)$ and s from the propensity score equation yields

$$\begin{aligned} P[Y = 1|X = x, S(j, x) = s] &= x\alpha_2 + \int_0^s F^{-1}_{M|X}(1 - s)ds \\ &= x\alpha_2 + K(s, x). \end{aligned} \tag{17}$$

Suppose that strong enough conditions on the data generating process hold so that $P[Y = 1|X = x, S(j, x) = s]$ is identified. Suppose in addition that $K(s, x)$ is not a linear function of x—i.e., that (i) there is at least one instrumental variable $\tilde{z}$ that moves the propensity score without being an exact linear function of x, (ii) $K(s, x)$ is a nonlinear function of x given s (or (iii) both of these conditions hold). Then $K(s, x)$ is not perfectly collinear with

[27] The assumption that the location shift in $F_{M|X}$ caused by x is linear can be weakened, so that the location shift is any known function $q(x)$. Then $q(x)$ will replace $x\alpha_0$ everywhere in the analysis above. Provided that $q(x)$ is not proportional to the integral of $F_0^{-1}(1 - s)$, all substantive aspects of the analysis are the same.

x, so $K(s,x)$ is the (nonempty) part of $P[Y=1|X=x, S(j,x)=s]$ that is orthogonal to x. That is enough to identify $K(s,x)$, and therefore $\delta(x,\tilde{z}_j)$, which is $\partial K(s(j,x)/\partial s$.

In sum, identification of $\delta(x,\tilde{z}_j)$ requires assumptions on $F_{M|X}$ and $\delta(\cdot)$ that are strong enough to identify $P[D=1|j,X=x]$ and $P[Y=1|X=x,S(j,x)=s]$, as well as that there be an instrumental variable or a useful nonlinearity in $K(s,x)$. The foregoing identification discussion may be usefully cast in terms of HV2005's MTE and HV1999's LIV parameters, to which I turn next.

6.5 Representing the stop-and-search model via MTE and LIV

In this section I show how to represent the stop-and-search behavioral model with the generalized Roy model discussed in HV1999 and HV2005.[28] Let Y_1 be an indicator variable that equals 1 if a stopped driver is carrying contraband and 0 otherwise. Let $Y_0=0$. Let D be an inidcator variable that equals 1 if a stopped driver is searched, and let $Y=DY_1+(1-D)Y_0=DY_1$ (given that $Y_0=0$). Because I have assumed that a search will always turn up any contraband that's present, Y equals 1 for stopped drivers who are searched and found with contraband and equals 0 for all other stopped drivers—i.e., drivers who are not searched and drivers who are searched but are not carrying contraband. Let $\mu_D(z)\equiv P[D=1|Z=z]$. Leaving aside some technical conditions, the representation of the stop-and-search behavioral model in generalized Roy model terms is

$$\begin{aligned} Y_0 &= 0 \\ Y_1 &= 1[1-U_1\geq 0] \\ D &= 1[U_D\leq \mu_D(Z)],\ U_D\sim \text{Unif(0,1)} \\ U_1 &= F_{M|X}(U_1^*|X=x)+U_D,\ U_1^*\sim \text{Unif(0,1)} \\ \tilde{Z} &\perp U_D|X, \end{aligned} \tag{18}$$

with U_D and U_1^* independent of each other and everything else in the structure.

Because $Y_0=U_0=0$ in the present context, we have $E[Y|X=x,P(Z)=p]=pE[Y_1|X=x]+pE[U_1|X=x,D=1]$. We have $E[Y|X=x,P(Z)=p]=pE[Y_1|X=x,P(Z)=p,D=1]=pE[Y_1|X=x,D=1]$, with the last equality following because Y_1 is independent of Z given X. Using the fact that $D=1$ and $U_D\leq p$ are the same event, we have $E[Y|X=x,P(Z)=p]=\int_0^p E[Y_1|X=x,U_D=u]du$. Now, $E[Y_1|X=x,U_D=u]=P(F_{M|X}(U_1^*|X=x)<1-u)=P(U_1^*<F_{M|X}^{-1}(1-u|X=x))$. Notice that $F_{M|X}^{-1}(1-u|X=x)$ is the search distaste that an officer facing a stop with characteristics $X=x$ must have to induce a search rate of u, i.e. $t_x(u)$.[29] Because U_1^* is uniform on the unit interval, $P(U_1^*<t_x(u))=t_x(u)$. Putting this all together yields

[28] See also CITES–Brinch et al., Cornellissen et al., etc.

[29] Here I use notation analogous to that developed in (8), with the x subscript indicating conditioning on characteristics $X=x$.

$$E[Y|X = x, P(Z) = p] = \int_0^p t_x(u)du. \tag{19}$$

By definition, the derivative of the left hand side of (19) with respect to p is Heckman & Vytlacil's (1999) local IV parameter. Heckman & Vytlacil (2005) show that this is also the MTE. Because Y is the unconditional hit dummy in the stop-and-search behavioral model, it is also true that $E[Y|X = x, P(Z) = p] = \upsilon(t_x(p))$, i.e., the unconditional hit rate function evaluated at the search distaste of an officer who would search at rate p. We saw in section 6.1 that the derivative of this function with respect to the search rate is $t_x(p)$. Differentiating the right hand side of (19) with respect to p also yields the same result: $t_x(p)$. This proves that the MTE in the generalized Roy model is the search distaste of the officer with search propensity p. This is just as Arnold et al. (Forthcoming$b$) state, albeit in different notation and with less detail, in their study of bail and racial discrimination. One helpful consequence is that all the machinery developed in the literature for estimation of MTE may be used here.[30]

One final observation involves the case when (11) defines the model's empirical structure, with $g = F_0$ as in Example 2. When g and F_0 are linear, the model yields two equations that are linear in all parameters. In the general nonlinear "double-F_0" case, it can be shown that although the propensity score equation has a simple form, the equation for $Y_1|D = 1$ involves the integral of a nonlinear function of parameters and the propensity score. Thus this equation will usually need to be estimated with semi- or nonparametric methods.

7 Estimates of Search Distastes

TO BE ADDED.

8 Conclusion

TO BE ADDED.

[30]See Appendix A for a description of the MTE estimation method used by ADY.

Appendix A ADY's Estimation Method

According to the online appendix to their paper, Arnold, Dobbie & Yang (Forthcoming*a*, pp. 34-35), ADY did the following to estimate the MTE.[31]

1. ADY residualized their outcome, selection, and instrumental variables with respect to covariates X by constructing $\dot{Y} \equiv M_X Y$, $\dot{\tilde{Z}} \equiv M_X \tilde{Z}$, and $\dot{D} \equiv M_X D$, where $M_X = I - P_X$ is the residual-maker matrix for covariates X, given that $P_X = X(X'X)^{-1}X'$ is the projection matrix for X.

2. Next, ADY construct $\widehat{p}$, which is the vector of fitted values from a projection of $\dot{D}$ onto $\dot{\tilde{Z}}$: $\widehat{p} = P_{\dot{\tilde{Z}}} \dot{D}$.

3. For observation i, ADY estimate $E[\dot{Y}_i | \widehat{p} = \widehat{p}_i]$ using a local quadratic estimator; call this $\widehat{\dot{Y}}(\widehat{p}_i)$.

4. ADY estimate $dE[Y_i | X = x_i, P(Z) = p_i]/dp_i$ using $\widehat{\dot{Y}}(\widehat{p}_i) - \widehat{\dot{Y}}(\widehat{p}'_i)$ for some value $\widehat{p}'_i$ that is close to $\widehat{p}_i$. These are the estimates graphed in Figure II of ADY's main text and discussed throughout the paper.

This procedure differs from (say) the approach in Heckman, Urzua & Vytlacil (2006) in a couple of ways. First, as ADY recognize, their $\widehat{p}$ is not a propensity score. Under their assumption that the first stage is linear in X and $\tilde{Z}$, $D = XB + \tilde{Z}\Gamma + W$ with $E[W|X, \tilde{Z}] = 0$, so the propensity score for $(X, \tilde{Z}) = (x, \tilde{z})$ is $p(x, \tilde{z}) = xB + \tilde{z}\Gamma$. This can be consistently estimated using the OLS estimate of the regression of D onto $(X, \tilde{Z})$. compare ADY's $\widehat{p}$, which is $P_{\dot{\tilde{Z}}} = \tilde{Z}\widehat{\Gamma} - P_X \tilde{Z}\widehat{\Gamma}$, where hats denote the OLS estimators of the parameter Γ. This estimates the part of the propensity score that is both (i) captured by variation in their judge-leniency instrument and (ii) orthogonal to variation in X covariates. Even in large samples, this will not generally equal the actual propensity score except by happenstance.[32] Under ADY's linearity assumptions, this deviation of $\widehat{p}$ from the propensity score is unproblematic, because X is implicitly included in all their subsequent specifications, and the actual propensity score under linearity is $\widehat{p} + P_X Z\widehat{\Gamma} + X\widehat{B}$. By construction, $\widehat{p}$ is orthogonal to the two latter terms. Thus the only effect of ADY's use of $\widehat{p}$ rather than a consistent estimator for the propensity score is to change the coefficients on the X covariates.

A second difference from the method in Heckman et al. (2006) involves the specification of the MTE implied by ADY's approach to estimating $E[Y|X = x, P(Z) = p]$. Their specification has no interaction between the propensity score and X. That implies that the local IV parameter does not vary with X. This is unusual in the MTE literature.

[31] I verified my understanding of ADY's procedure with the authors via email correspondence.

[32] Although ADY describe $\widehat{p}$ as a propensity score, their discussion indicates that they regard the orthogonality of $\widehat{p}$ as a feature, because "the treatment effects we are interested in are identified by variation in judge leniency by definition"; see Arnold et al. (Forthcoming*a*, pp. 34-35, footnote 2).

Consider the example in (11) with $g = F_0$ both being linear functions. Then we have

$$E\left(Y|X=x, S(j,X)=s\right) = \\ x\alpha_2 + (x\alpha_0)s + c\left[s - \frac{1}{2}s^2\right], \tag{20}$$

where c is some constant. ADY's step 3 can be viewed as estimating this equation. By excluding interactions of X and s from their estimation, ADY implicitly impose $\alpha_0 = 0$. The substantive content of that assumption is that $F_{M|X}(\delta|X = x) = F_0(\delta)$. This means there is no variation in the defendant misconduct distribution due to X. This is a strong assumption, and not one that ADY discuss.

Third, consider estimating (20) with $\alpha_0 = 0$ imposed. In my notation, the approach taken in Appendix B.2 of Heckman et al. (2006) is to estimate local regressions of Y on s and X on s, calculate the residuals, and then calculate the OLS coefficient of the residualized Y on the residualized X. This yields an estimate of α_2. Then Heckman et al. (2006) calculate $Y - x\widehat{\alpha}_2$ and use local linear regression of this quantity on the propensity score s to recover what in my notation is $\int_0^s F_0^{-1}(1-s)ds$. [33] But ADY use a different approach than this. Steps 1 and 2 of their procedure, described above, do not partial out the variation in s from the Y and X variables. Rather, these steps partial out the variation in X from the Y and s variables. If there were no nonlinear term in s in (20), then this procedure might be valid. But with a nonlinear term, the procedure appears to break down, because the nonlinear function of a residualized variable is not the same as the residualized nonlinear function of the variable. Thus it is not immediately clear what ADY's step-3 local quadratic estimator actually estimates.

Notice that if there were no quadratic term in (20), then the problem in the previous paragraph. We would be left with a constant MTE. Although this is consistent with ADY's Figure II, at least for black defendants, it means that the officer-level distaste parameter is invariant to everything. But "everything" includes both X ad the instrumental variable $\tilde{z}_j$ (in ADY's terms, the judge leniency variable). That would appear to leave us with a model in which no observed variable is associated with either the defendant misconduct distribution or judge tastes. That is contrary to both intuition and ADY's other estimates. Accordingly there must be some nonlinearity in the way s enters (20).

References

Antonovics, K. & Knight, B. (2009), 'A new look at racial profiling: Evidence from the boston police department', *Review of Economics and Statistics* **91**, 163–77.

[33] It happens that under the assumption that the propensity score equation is linear, as ADY treat it, this function is quadratic, i.e., $c[s - \frac{1}{2}s^2]$ (see (20)). Thus ADY's linearity assumption on the propensity score would be strong enough to allow parametric estimation of (20) with a quadratic in the propensity score included (see, e.g., **?** for a general discussion of such a "global polynomial" approach).

Anwar, S. & Fang, H. (2006), 'An alternative test of racial prejudice in motor vehicle searches: Theory and evidence', *American Economic Review* pp. 127–151.

Arnold, D., Dobbie, W. & Yang, C. S. (Forthcoming*a*), 'Online appendix: Racial bias in bail decisions', *Quarterly Journal of Economics* .

Arnold, D., Dobbie, W. & Yang, C. S. (Forthcoming*b*), 'Racial bias in bail decisions', *Quarterly Journal of Economics* .
URL: *Pre-publication version at https:drive.google.comopen?id1Fm5RgydjvNrIlpaX8X38ugBAIQIgCn*

Ayres, I. (2002), 'Outcome tests of racial disparities in police practices', *Justice Research and Policy* **4**.

Ayres, I. & Waldfogel, J. (1994), 'A market test for race discrimination in bail setting', *Stanford Law Review* **46**.

Becker, G. S. (1957), *The economics of discrimination*, University of Chicago Press, Chicago.

Bjerk, D. (2007), 'Racial profiling, statistical discrimination, and the effect of a colorblind policy on the crime rate', *Journal of Public Economic Theory* **9**, 543–67.

Dharmapala, D. & Ross, S. L. (2004), 'Racial bias in motor vehicle searches: Additional theory and evidence', *The B.E. Journal of Economic Analysis & Policy* **0**(1).

Grogger, J. & Ridgeway, G. (2006), 'Testing for racial profiling in traffic stops from behind a veil of darkness', *Journal of the American Statistical Association* **101**, 878–87.

Heckman, J. J., Urzua, S. & Vytlacil, E. (2006), 'Understanding instrumental variables in models with essential heterogeneity', *Review of Economics and Statistics* **88**(3), 389–432.

Heckman, J. J. & Vytlacil, E. (1999), 'Local instrumental variables and latent variable models for identifying and bounding treatment effects', *Proceedings of the National Academy of Sciences* **96**, 4730–34.

Heckman, J. J. & Vytlacil, E. (2005), 'Structural equations, treatment effects, and econometric policy evaluation', *Econometrica* **73**(3), 669–738.

Horrace, W. C. & Rohlin, S. M. (2016), 'How dark is dark? bright lights, big city, racial profiling', *Review of Economics and Statistics* **98**(2), 226–32.

Imbens, G. W. & Angrist, J. D. (1994), 'Identification and estimation of local average treatment effects', *Econometrica* **62**(2), 467 – 75.

Kalinowski, J., Ross, S. L. & Ross, M. B. (2017), Endogenous driving behavior in veil of darkness tests for racial profiling, Working Paper 2017-03, University of Connecticut Department of Economics Working Paper.
URL: *http://web2.uconn.edu/economics/working/2017-03.pdf*

Knowles, J., Persico, N. & Todd, P. (2001), 'Racial bias in motor vehicle searches: Theory and evidence', *Journal of Political Economy* **109**, 203–29.

Persico, N. G. & Todd, P. (2006), 'Generalising the hit rates test for racial bias in law enforcement, with an application to vehicle searches in wichita', *Economic Journal* **116**(515), F351–F367.

Pierson, E., Simoiu, C., Overgoor, J., Corbett-Davies, S., Ramachandran, V., Phillips, C. & Goel, S. (2017), A large-scale analysis of racial disparities in police stops across the united states, Working paper, Stanford University.
URL: *https://5harad.com/papers/traffic-stops.pdf*

Sanga, S. (2009), 'Reconsidering racial bias in motor vehicle searches: Theory and evidence', *Journal of Political Economy* **117**(6), 1155–59.

Simoiu, C., Corbett-Davies, S. & Goel, S. (2017), 'The problem of infra-marginality in outcome tests for discrimination', *The Annals of Applied Statistics* (3), 1193–1216.

Vytlacil, E. (2002), 'Independence, monotonicity, and latent index models: An equivalence result', *Econometrica* **70**, 331–41.

Figure 1: Kernel density estimates for number of drivers stopped by officer (individual-level data)

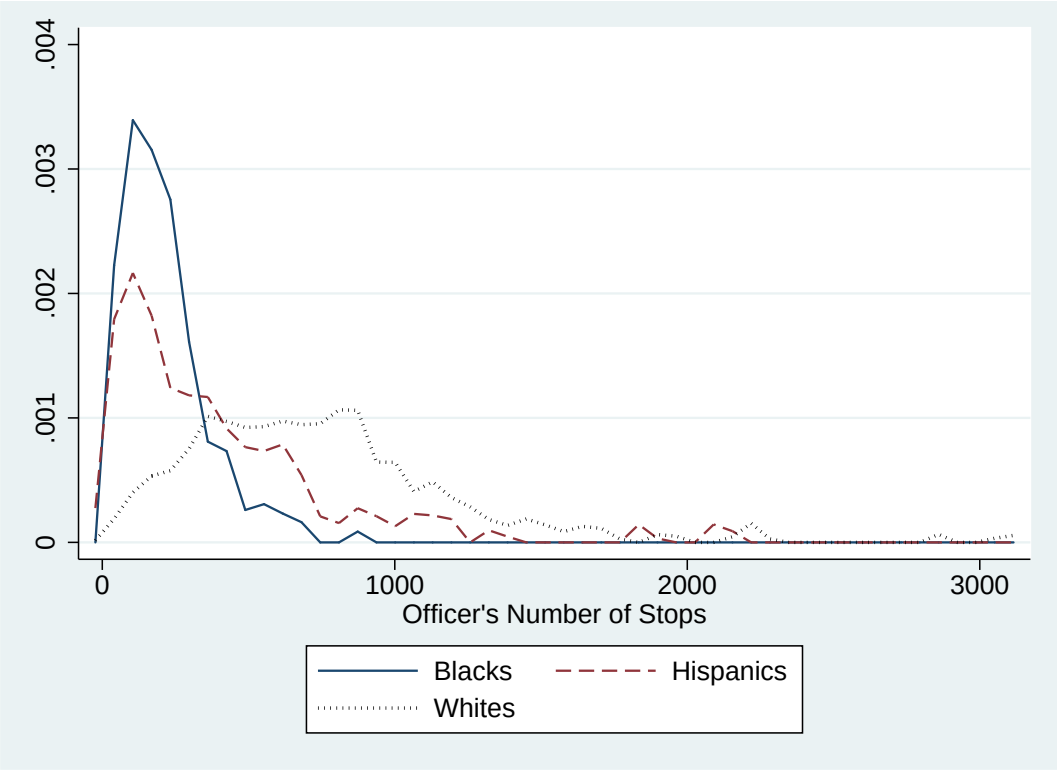


Figure 2: Kernel density estimates for leave-one-out officer-level search rate variable (among those searched)

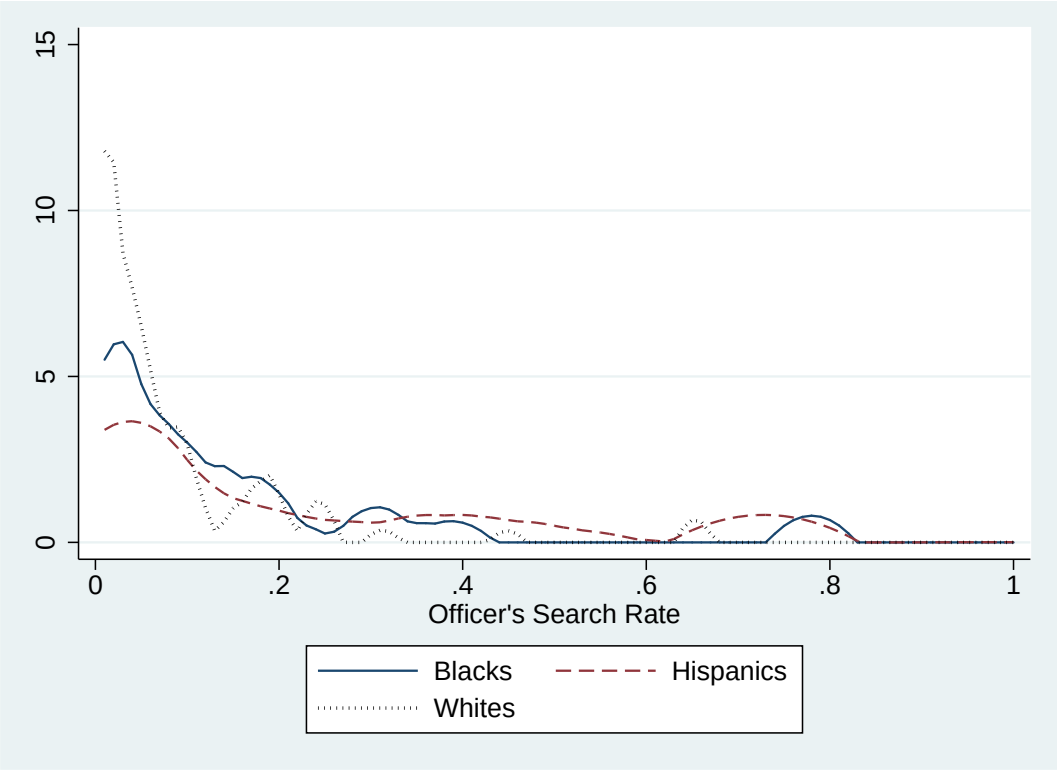


Table 1: Means for variables used in analysis (Anwar & Fang Florida data)
(generating time: Mon Jul 23 11:49:05 2018 from file *summary-stats.ara*, table 1.)

	Full sample	Black	Hispanic	White
Search rate				
Search was conducted	0.010	0.013	0.013	0.008
Hit rates				
Among only those searched	0.210	0.209	0.115	0.251
Among all drivers	0.002	0.003	0.002	0.002
Driver characteristics				
Male	0.696	0.672	0.776	0.681
Black	0.162	1.000	0.000	0.000
Hispanic	0.173	0.000	1.000	0.000
White	0.665	0.000	0.000	1.000
Officer characteristics				
Black	0.160	0.208	0.196	0.139
Hispanic	0.114	0.101	0.230	0.088
White	0.726	0.691	0.574	0.774
Male	0.907	0.910	0.895	0.909
Stop/car characteristics				
Stop was between 6am and 6pm	0.697	0.670	0.668	0.711
Out-of-state registration	0.101	0.101	0.051	0.114
At least one passenger	0.402	0.448	0.425	0.385
N	906,339	146,860	156,784	602,695

Note: Data from same file used by Anwar & Fang. All variables are binary.

Table 2: Florida Coefficients From OLS Estimation of Relationship Between Conditional Hit Rate and Leave-One-Out Officer-Level Search Rate (using county dummies dummies to control for location)
(generating time: Sun Jul 22 07:47:09 2018 from file *hitrate-on-officer-search-variable.ara*, table 1.)

	Black	Hispanic	White
Officer-level search rate	-0.140	-0.185	-0.196
	(0.070)	(0.042)	(0.102)
Officer characteristics			
Black	-0.009	0.071	0.086
	(0.051)	(0.059)	(0.069)
Hispanic	-0.026	0.085	0.013
	(0.038)	(0.039)	(0.037)
Officer male	0.059	0.020	0.030
	(0.052)	(0.045)	(0.031)
Driver/stop characteristics			
Driver male	-0.004	0.008	0.026
	(0.027)	(0.019)	(0.016)
Stop between 6am and 6pm	-0.029	0.003	0.029
	(0.020)	(0.019)	(0.017)
Out-of-state license plates	-0.027	-0.032	-0.033
	(0.030)	(0.019)	(0.017)
At least one passenger	0.048	0.042	0.083
	(0.018)	(0.014)	(0.013)
N	1,979	2,094	4,897

footer *Note*: Data from same file used by Anwar & Fang. All variables are binary. Numbers in parentheses are standard errors estimated allowing for clustering at the officer level. Estimation results for coefficients on county dummies not reported.

Table 3: Florida Coefficients From OLS Estimation of Relationship Between Conditional Hit Rate and Leave-One-Out Officer-Level Search Rate (using five district to control for location)
(generating time: Sun Jul 22 07:47:10 2018 from file *hitrate-on-officer-search-variable.ara*, table 2.)

	Black	Hispanic	White
Officer-level search rate	-0.173	-0.194	-0.338
	(0.051)	(0.044)	(0.079)
Officer characteristics			
Black	0.040	0.065	0.097
	(0.050)	(0.048)	(0.067)
Hispanic	-0.066	0.089	0.016
	(0.032)	(0.047)	(0.042)
Officer male	0.025	0.008	0.027
	(0.046)	(0.036)	(0.040)
Driver/stop characteristics			
Driver male	-0.003	0.002	0.039
	(0.026)	(0.023)	(0.017)
Stop between 6am and 6pm	-0.019	0.015	0.029
	(0.020)	(0.018)	(0.017)
Out-of-state license plates	-0.028	-0.055	-0.045
	(0.030)	(0.019)	(0.019)
At least one passenger	0.061	0.047	0.083
	(0.018)	(0.015)	(0.013)
N	1,979	2,094	4,897

footer *Note*: Data from same file used by Anwar & Fang. All variables are binary. Numbers in parentheses are standard errors estimated allowing for clustering at the officer level. Estimation results for coefficients on district dummies not reported.

Table 4: Means for variables used in analysis (Harris County, Texas)
(generating time: Mon Jul 23 12:37:00 2018 from file *summary-stats-harris.ara*, table 1.)

	Full sample	Black	Hispanic	White
Search rate				
Search was conducted	0.016	0.022	0.015	0.014
Hit rates				
Among only those searched	0.334	0.349	0.303	0.340
Among all drivers	0.005	0.008	0.005	0.005
Driver characteristics				
Male	0.657	0.606	0.710	0.650
Black	0.256	1.000	0.000	0.000
Hispanic	0.266	0.000	1.000	0.000
White	0.434	0.000	0.000	1.000
Stop characteristics				
Between 6am and 6pm	0.556	0.536	0.522	0.595

Note: Data from Stanford Open Policing Project Texas file.

Table 5: Harris County Coefficients From OLS Estimation of Relationship Between Conditional Hit Rate and Leave-One-Out Officer-Level Search Rate (including fine-grained-location dummies)
(generating time: Mon Jul 23 12:48:37 2018 from file *hitrate-on-officer-search-variable-harris.ara*, table 1.)

	Black	Hispanic	White
Officer-level search rate	-0.087	0.802	1.891
	(0.557)	(0.255)	(0.648)
Driver male	0.047	0.052	-0.022
	(0.024)	(0.035)	(0.022)
Stop between 6am and 6pm	-0.051	-0.007	-0.040
	(0.021)	(0.022)	(0.020)
N	3,175	2,149	3,221

footer *Note*: Data from Stanford Open Policing Project Texas file. All variables are binary except officer-level search rate. Estimation results not reported for coefficients on year dummies and fine-grained-location dummies. Numbers in parentheses are standard errors estimated allowing for clustering at the officer level.