“Whose help is on the way?”
The importance of individual police officers in law enforcement outcomes

Emily Weisburst*

June 2019
(First Version: July 2017)

Link to Download Latest Version

Abstract

Police discretion has large potential consequences for civilian trust, public safety and individuals that directly interact with the police. However, little is known about how much police discretion actually contributes to law enforcement outcomes and whether police officers are fair and impartial in their application of the law. In this paper, I show that the likelihood of an arrest is not only a function of incident type, timing, geography and urgency, but also critically depends on the identity of the police officer who responds to a call for service. The analysis examines detailed data on more than 1,500 police officers responding to nearly 2 million calls for service from Dallas, Texas. Officers vary widely in their arrest behavior, with a 1 standard deviation increase in an officer's propensity to arrest resulting in a 37% increase in the likelihood of arrest. High and low arrest officers face similar crime offending environments; however, high arrest officers are more likely to use physical force during an arrest, arrest individuals for less severe crimes and book arrested individuals in jail for very low-level offenses. Additionally, despite documenting large racial disparities in arrests, I fail to find conclusive evidence that officer arrest behavior is characterized by taste-based racial bias in this setting.

JEL Codes: K40, K42, J45

Keywords: Police Discretion and Decision-Making, Racial Bias in Policing, Economics of Crime

*Assistant Professor, Luskin School of Public Affairs, University of California, Los Angeles. Contact address: Department of Public Policy, UCLA Luskin School of Public Affairs, 337 Charles E. Young Dr E., Los Angeles, CA 90095. Email: weisburst@ucla.edu. I am grateful to Professors Sandra Black, Jason Abrevaya, Leigh Linden, Becky Pettit, Rich Murphy, Steve Trejo and Brendan Kline for their helpful comments and feedback. I also thank members of the Dallas Police Department that shared information about their operations, protocols and data practices for this project. This project received funding from the University of Texas Continuing Graduate Fellowship. The views presented in this paper are my own and do not necessarily reflect the views of the Dallas Police Department.

1 Introduction

In recent years, public debate about the fairness of police practices has escalated. Recent survey evidence finds that less than 30% of individuals in high-crime and low-income areas believe that police "make fair and impartial decisions in the cases they deal with" or that police "make decisions based on the law and not their personal opinions or beliefs" (La Vigne et al., 2017). In 2015, American confidence in police officers reached its lowest point in more than 20 years, driven by high profile police use-of-force incidents and shootings (Jones, 2015). In the ensuing debate, pundits have made conflicting claims about the nature of police behavior during these high profile incidents, sometimes asserting that "any officer would have responded in the same way" and at other times claiming that the events are "isolated incidents attributable to 'bad actors' that do not reflect the rest of a department."[1]

Police discretion has large potential consequences for civilian trust, public safety and individuals that directly interact with the police. However, on a basic level, there is scant evidence of whether officer decisions actually matter to the outcomes of police interactions after considering the context of an incident. If there are differences in police officer behavior, how large are these differences? Further, how do differences in officer behavior relate to characteristics of police outcomes and racial bias?

Managers in police departments face a similar principal-agent problem as managers in firms; they are impacted by the actions of individual officers but cannot fully control officer behavior given limited resources. Understanding the trade-offs between alternative monitoring policies is an important area of study, particularly in the high stakes setting of policing. This paper measures the scope of individual police officers to impact law enforcement outcomes, a necessary first step to clarifying these trade-offs.

Police officers often operate in the field, alone or in small groups and have substantial legal latitude in their conduct and response to different situations. At the same time, police departments are increasingly incorporating technology and data to standardize operations,

[1]This phrasing is not directly attributed to any single pundit or public figure. An example of the first argument can be found in opinion pieces through the organization *Blue Lives Matter*, which was established as a reaction to the *Black Lives Matter* movement (BlueLivesMatter, 2016). The second argument was recently invoked by Attorney General Sessions as a reason to cease the Department of Justice's enforcement of consent decree agreements with police departments, which were established to address civil rights concerns related to law enforcement actions (Department of Justice, 2017).

potentially limiting the importance of individual officer decisions in police work.[2] The ability of police officers to invest effort differently across different types of offenses can be a productive form of discretion when police resources are limited and there is a trade-off between exerting effort on serious and non-serious crimes. However, within particular incident types, behavioral differences across officers are more likely to result from differences in officer skills, experience and preferences. This study is the first to estimate the degree and importance of police discretion across officers, conditional on incident context. The project also addresses potential implications of police discretion, by relating individual officer behavior to characteristics of police outcomes and patterns of racial bias.

I analyze police officer arrest decisions using a sample of nearly 2 million calls for service (or 9-1-1 calls) and over 1,500 police officers from the Dallas Police Department in Texas. I estimate individual officer arrest propensity, controlling for detailed information on the characteristics of calls, including call urgency and dispatch code, peer responders and time and geographic factors. Throughout the analysis, I pay particular attention to patterns of officer sorting to calls for service and conduct a number of robustness checks to verify that the observed dispersion in arrest behavior across officers is not due to selection.

There is substantial variability in police officer responses to similar offenses. I find that a 1 standard deviation increase in an officer's arrest propensity results in a 37% increase in the likelihood of arrest. The dispersion in predicted arrest outcomes across officers is comparable to the dispersion across geography and is approximately a third of the dispersion attributable to call type. Linking the officer arrest propensity estimates to officer demographics, I find that Black officers and more experienced officers have moderately lower arrest propensities.

Next, I investigate relationships between officer arrest propensity and characteristics of arrest outcomes. While high and low arrest officers face similar crime offending environments, high arrest officers are more likely to use physical force during an arrest, make arrests

[2] Criminologists have long noted that police work is characterized by discretion, with researchers adopting a broad definition of discretion that encompasses variation in work-related decisions, interpretation and implementation of the law and the use of extra-legal factors, such as suspect race, in decision-making (Nickels, 2007; Frydl and Skogan, 2004; Mastrofski, 2004). Technological advancements have increased police reliance on data for surveillance of suspects, tracking and monitoring of police employee activities, automation of reporting and focusing patrol on areas with high offending rates, or crime "hot spots." Advancements in technology have the potential to exacerbate differences in police treatment of civilians, or could serve to reduce police discretion (Brayne, 2017; Joh, 2016).

for less severe crimes and book individuals in jail for very low-level offenses. These results suggest that high arrest officers may have a lower severity or evidence threshold for making arrests and applying sanctions.

I fail to find conclusive evidence of taste-based racial bias among officers, despite the fact that I document large racial disparities in arrest outcomes in Dallas, where Black civilians account for 50% of arrests but only 25% of the population (Census, 2015). I adapt a relative test of taste-based racial bias from Anwar and Fang (2006) to the setting of calls for service, where not all incidents may have a potential suspect. I find that high arrest officers are more likely to arrest an individual of any race than low arrest officers and that Black officers are less likely to arrest an individual of any race than Hispanic or White officers. This evidence is consistent with large officer differences the net benefits of making arrests of individuals from any race group, rather than differences in racial bias across officers.

In economics, the research on police decision-making has largely focused on measuring racial bias in police traffic stops (e.g. Horrace and Rohlin, 2016; Anbarci and Lee, 2014; Antonovics and Knight, 2009; Gelman et al., 2007; Anwar and Fang, 2006; Grogger and Ridgeway, 2006; Knowles et al., 2001). New work by Fryer (2016) extends this literature to officer decisions to use violent force. Collectively, this literature has found mixed evidence that police officers exhibit taste-based preferences for racial discrimination, with results that vary by the study setting and the test used to detect racial bias. While the literature has frequently exploited aggregate officer demographic characteristics, nearly all of the work in this space does not incorporate officer identity in measuring racial bias. This paper complements new work by Goncalves and Mello (2017), that measures individual officer effects in a racial bias test applied to officer decisions to issue speeding tickets.

In addition to providing the first estimate of officer-level police discretion, this paper makes several other contributions. First, the analysis in this study uses police responses to 9-1-1 calls for service, a setting that researchers have not exploited to study police decision-making.[3] I am able to explicitly measure the importance of officers versus incident context because the incident setting is given at the time of the police response.

[3]To the best of my knowledge, this is the first study that uses high frequency complainant offense reports to study police officer behavior and decisions.

With the exception of West (2015), who studies racial bias of state troopers who are randomly dispatched to motor vehicle accidents, researchers in economics have typically restricted their attention to interactions that are initiated by police officers, such as general traffic stops, stop and frisk interviews and speeding tickets. Importantly, these interactions are a choice variable of the police officers involved. A growing body of research finds that race can also be a factor in police decisions to make traffic (or pedestrian) stops and that studies that focus only on the outcomes of traffic stops neglect to consider police discretion that contributes to sample selection (Horrace and Rohlin, 2016; Gelman et al., 2007; Grogger and Ridgeway, 2006). A major advantage of using call for service data to study police discretion is that each observed call response is originally initiated by a complainant and not by a police officer.

Second, I am able to link variation in officer discretion to characteristics of arrest outcomes. I draw on multiple data sources to construct several measures of arrest characteristics and outcomes, covering officer use-of-force, arrest charge severity, jail admission and court dismissal or conviction outcomes. These measures of arrest characteristics inform our understanding of multiple dimensions of officer arrest behavior and provide insight about the attributes of marginal arrests.

An additional contribution of this study is the variety of policing interactions that I examine. The analysis sample covers a diverse cross-section of police work, allowing examination of responses to different types of offenses. Detailed demographic information on officers, arrestees and civilian complainants also provide rich controls in the model and enable tests of racial bias.

2 Institutional Background and Description of Data

2.1 *Dallas Police Department Data*

The setting for this study is the Dallas Police Department in Texas. Dallas is a large and diverse urban center, with over 1.2 million residents and a population that is 42% Hispanic, 24% Black and 29% White (Census, 2015).[4] Crime rates in the city of Dallas are similar to other cities of its size in the U.S., with 694 violent crimes and 3,440 property crimes per 100,000 residents in 2015 (FBI, 2015).

In recent years, the Dallas Police Department (DPD) has enacted several police reforms, including increasing officer training in de-escalation techniques and racial bias, employing body cameras and firing some of its most poorly performing officers (Haugh, 2016; Tsiaperas, 2016). In 2015, DPD joined the Obama Administration's Police Data Initiative and released a number of data sets on its operations.

This project uses administrative DPD data covering dispatched calls for service (9-1-1 calls), records of persons arrested and the names and badge numbers of responding officers between June 2014 and October 2018. This data is supplemented with information on individuals suspected of offenses, non-shooting use-of-force incidents, jail booking records and court records for the analysis of measures of policing quality and the test of taste-based racial bias. I construct the primary arrest outcome using a liberal definition of arrests, coding an offense as having an arrest if any of the DPD data files obtained for the study indicate that an arrest was made. Additionally, I merge the DPD data sets with demographic information on police officers also obtained through an open records request to the city of Dallas. The officer data includes officer race, gender, salary and job title.[5]

2.2 *Protocols for Dispatch and Offense Responses*

When a civilian calls DPD for police assistance, they are connected to a 9-1-1 call-taker. The call-taker creates an active call report that summarizes important facts related to the incident, including location and relevant descriptions of the events. Active call reports also include a dispatch code that categorizes the incident type. Given a set of open active calls, DPD dispatchers work with police officers to assign available officers to incidents. Calls are dispatched according to their priority, or their level of severity and urgency. When there is a long call queue, responses to low priority calls are postponed until more serious calls

[4] I capitalize Black, Hispanic and White for stylistic consistency throughout the paper.

[5] See the Online Data Appendix A5 for more detail on the data sets and for information on data construction and cleaning.

have been resolved. The pool of available officers when a call is received depends on patrol responses to other incidents at the time. (Figure A1 depicts the response process in Dallas).

Patrol officers are the primary responders to calls for service. Officers are assigned to work in 1 of 7 police divisions in the city for 8-hour shifts, or watches, from 12am-8am, 8am-4pm and 4pm-12am.[6] Regular patrol shift schedules are set once a year, based on the seniority of officers.[7] With rare exceptions, calls are strictly assigned to patrol officers who work within the geographic police division where the call incident occurred.

Officers typically conduct patrol in police cars, alone or in pairs. At the beginning of each shift, officers may choose to patrol with another officer, depending on the number of cars available for that shift. Each car is considered an "element" that can be dispatched to an incident. Paired officers respond to all calls together throughout a shift.

If more than one patrol element is available to respond to an incident at the time of dispatch, dispatchers consider a number of factors in their assignment of available officers. More serious incidents may require or benefit from a response by multiple officers or "elements." Additionally, officers who are geographically close to an incident are more likely to be dispatched to the incident, especially if the call is urgent. At the same time, depending on availability, officers may volunteer to take particular calls as they are posted, a potential source of selection.

This project focuses on the first group of officers to be dispatched to a call, based on the call dispatch time stamp. For computational purposes, the analysis sample is also restricted to exclude calls dispatched more than 90 minutes after the call is received and responses with more than 4 dispatched officers.

The estimates in the study will impacted by selection bias if officers choose to respond to calls based on incident characteristics that are unobservable. To address this concern, I conduct a series of tests to verify that selection does not affect the empirical estimates (see Section 4.3).

When the assigned patrol element arrives at the scene of the incident, the responding

[6]The three 8-hour shifts listed are approximate, in practice some officers work 10 hour shifts and other officers have start and end times that are slightly staggered across police shifts.

[7]Depending on the needs of the department, officers may choose to work overtime patrol shifts outside of their regular shift schedules, though these shifts are also set in advance, typically a month or a week prior.

officer(s) determines if an offense occurred, gather information, investigates the scene and assists the complainant or victim. If an officer determines that an offense occurred, the officer submits an offense report to a staff reviewer at DPD who examines the report for completeness. After this initial review, the offense may be assigned to a detective in an investigative unit based on the offense type. The assigned detective will then pursue additional investigation of the offense if warranted.

Over the course of a call response, officers may identify a suspect and/or make an arrest. Alternatively, an arrest may be made at a later date after a detective assumes responsibility for a follow-up investigation of the offense. 3.5% of calls for service result in an arrest. Individual responding officers have the ability to influence arrests directly, by making the decision to apprehend an individual at the scene of the incident, or indirectly, by laying the groundwork for an investigation through gathering information for the initial offense report. In practice, most arrests do not involve a prolonged follow-up investigation and the responding officer is typically involved in the arrest.[8]

3 Empirical Model

I use a predictive model of arrests to estimate each officer's arrest propensity and measure the dispersion in this propensity across officers.

As a first step, I estimate the following linear probability model,

$$Arrest_{ikgt} \quad = \quad \theta_i + \theta_{-i} + \pi X_{kt} + \delta_{dt} + \phi_g + \varepsilon_{ikgt}$$

where i indexes the responding officer, $-i$ indexes a co-responding officer group effect (if other officers are present), k indexes the offense, d indexes geographic police divisions, g indexes geographic police beat location and t indexes time. The outcome $Arrest_{ikgt}$ is the primary focus of the analysis and denotes whether an arrest was made in association with the offense. X_{kt} are a set of incident specific characteristics, including 22 aggregated dispatch

[8] Of the data with information on arrest dates (89% of arrests), 98% of arrests occur within a day of the incident offense. When there is information on the arresting officer (63% of arrests), 84% of arrests involve the original responding officer.

codes and indicators for hour of day. The model also controls for the *urgency or severity* of the call, defined as the number of minutes that pass between when a call occurs and the time of dispatch (entered as a linear and quadratic term). Additionally, X_{kt} includes the proportion of officers available to be dispatched (relative to those working a shift) at the time of each call event.

ϕ_g are indicators for police beat locations and control for time-constant differences in arrest patterns across geography. There are 234 beats in Dallas and each is fully contained within 1 of the 7 police divisions in the city. δ_{dt} are shift indicators that capture time-varying location-specific arrest patterns that are associated with specific shift assignments. These variables are Police Division*Day-of-the-Week*8-hour Shift*Month*Year fixed effects. To increase power, the baseline model does not include a separate indicator for each individual shift, but rather aggregates shifts into month by year groups.[9] For example, the four Tuesday evening shifts in the Central Division are grouped in January 2016.

θ_i measures the time-invariant or permanent arrest propensity of officer i. Given the numerous controls in the empirical model, θ_i represents an officer specific effect that is measured within dispatch type, shift cell and geographic location. In cases where there is more than one responding officer, I include a control for the group of other responding officers, θ_{-i}. Observations with multiple responders are duplicated, so that each officer gets a record of participating in the offense response through the θ_i term. This procedure allows measurement of individual officer effects net of the effects of co-responders, using a model similar to prior work on peer effects in production (e.g. Silver, 2016; Mas and Moretti, 2009). In this way, the specification addresses omitted variable bias related to police officer decisions to pair with other officers as well as potential direct effects attributable to the co-responder. I restrict the sample to observations where the officers responding have at least 1,000 call records to improve precision in the estimation of θ_i.[10] [11]

[9] A version of the model with controls for individual 8-hour shifts is estimated as a robustness check in column (4) of Appendix Table A2.

[10] This restriction excludes 7% of officer-call observations in the sample, but allows estimation of fixed effects to be based on a reasonable number of observations per officer. 14.4% of co-responder offense responses include only records for some of the responders given this restriction (all co-responders are always considered in the co-responder group fixed effects controls). Further limiting the sample to exclude these "one-sided" observations does not affect the results.

[11] I use an algorithm developed in Correia (2016) to estimate the large number of fixed effects indirectly

Using this model, I calculate the dispersion in officer-level permanent arrest propensity as the standard deviation of the distribution of θ_i across officers. In order to establish a conservative estimate of police officer dispersion, I adjust the estimates of θ_i terms using Empirical Bayes techniques.[12] Throughout this paper, I focus on results using these adjusted estimates and refer to these adjusted estimates as $\hat{\theta}_i$. The results are not an artifact of this adjustment and are qualitatively similar when unadjusted fixed effects from the first stage are used (see Section 4.4 for a discussion of alternate precision adjustments).

4 Police Discretion Results

4.1 *Summary Statistics*

Table 1 summarizes the analysis data. The first column covers the total administrative sample at the call-level, while the second column covers the analysis sample, which excludes calls dispatched over 90 minutes after the call was received and calls with more than 4 responding officers and restricts the sample to observations for officers with 1,000 or more call responses. The analysis sample includes over 1.9 million calls and over 3.4 million observations for 1,544 officers. The samples summarized in Tables 1 are very similar and this consistency suggests it is suitable to generalize results from the analysis sample.

On average, it takes 14.5 minutes for a patrol officer to be dispatched to an offense incident after a call is made, with a standard deviation of 20 minutes. The variation in this

through an iterative procedure that provides a point estimate for each fixed effect. In the base model, there are 3,206,609 observations, 1,544 first officer categories, 93,077 co-responder group categories, 7,644 shift categories and 234 beats (after excluding singletons). Rather than estimating the model's fixed effects by including indicator variables as controls in the model, this algorithm initializes each fixed effect within a group and then iterates the estimation until both the sum of squared residuals is minimized and the coefficient on each set of fixed effect terms is 1. This procedure is programmed in the STATA command `reghdfe` and is notable for its fast computation time. I estimate all sets of fixed effects in the model jointly in this way, or $\theta_i, \theta_{-i}, \delta_{dt}$ and ϕ_g.

[12] I calculate the adjusted estimates of θ_i using the following steps. First, I construct a composite residual term, $\hat{r}_{ikgt} = \hat{\theta}_i + \hat{\varepsilon}_{ikgt}$ and an average officer residual, $\bar{r}_i = \frac{1}{N_i}\sum_{N_i} \hat{r}_{ikgt}$. Next, I calculate the adjusted officer arrest propensity using the following transformation: $\hat{\theta}_i^{EB} = \sigma_A^2/(\sigma_A^2 + \frac{\sigma_{\varepsilon,i}^2}{N_i}) \cdot \bar{r}_i$. The value of $\sigma_A^2 = \sigma_r^2 - \sigma_\varepsilon^2$, where σ_r^2 is computed using the sample analog of the average squared composite residual and σ_ε^2 is the average squared within officer composite residual, each calculated from the first stage model. The "shrinkage factor," $\sigma_A^2/(\sigma_A^2 + \frac{\sigma_{\varepsilon,i}^2}{N_i})$, adjusts officer arrest propensity toward zero when the number of observations per officer, N_i, is small, or the variation in the officer effect, $\sigma_{\varepsilon,i}^2$, is large (see Appendix A3 for more detail).

dispatch time lag highlights the fact that dispatchers prioritize calls based on their severity and that officers cannot immediately respond to all incidents. On average, 28% of officers on a shift are unavailable, or responding to other calls, at the time of each call dispatched. The most common dispatch codes are for major disturbances and burglaries of motor vehicles and residences. At the time of dispatch, only a small number of offense incidents are designated as known violent offenses; robberies, criminal assaults, armed encounters and active shootings collectively comprise less than 5% of calls.

Approximately 3.5% of call responses result in an arrest. White officers and Black arrestees are over-represented relative to the population of Dallas. White patrol officers respond to 46% of offenses, while Black and Hispanic officers respond to 26% and 21% of offenses, respectively. Relative to the sample of offenses with demographic information for arrestees, 50% of offenses have a Black arrestee, 26% have a White arrestee and 23% have a Hispanic arrestee.[13] 7% of call responses involve a police officer in training, 1% involve a police sergeant and 14% involve a female officer in the analysis sample. Averaged across call responses, DPD patrol officers earn approximately $57,000 per year.

4.2 *Baseline Results*

Individual police officers vary substantially in their arrest behavior. Figure 1 shows the estimated distribution of officer effects, $\hat{\theta}_i$, calculated using the procedure described in Section 3. For each officer, $\hat{\theta}_i$ represents his/her permanent or time-invariant arrest propensity, conditional on time and geography controls, call characteristics and peer influence. This estimated distribution has a longer right tail, showing that a small number of officers have especially high arrest propensities.

Swapping an officer that has a low arrest propensity with one that has a high arrest propensity can critically change the outcome of a call response. A 1 standard deviation in $\hat{\theta}_i$ corresponds to 0.07 standard deviations in the total arrest outcome. In percentage terms, a 1 standard deviation increase in an officer's arrest propensity results in a 37% increase in the likelihood that a given offense results in an arrest. Further, moving from the 10th to 90th percentile in the officer distribution translates to a 84% increase in arrest probability.

[13] Demographic information is not available for 10% of arrests in the sample.

The dispersion in officer effects is meaningful when benchmarked to other dimensions of call response. Figure 2 displays estimates for a standard deviation change in different model components. Call type is the most important predictor of whether an arrest will result from a call response, as differences in dispatch call code translate to the largest differences in arrest outcomes. A 1 standard deviation increase in the dispatch code effect distribution leads to nearly 100% increase in the likelihood of arrest. The variation in officer effects is comparable to the variation in geographic police beats, a result that is striking given the fact that policing practices vary widely across neighborhoods with differing crime, income and racial compositions. The estimated variation in the urgency of calls and time dimensions of calls (month effects, day of week effects and hour of day effects) is smaller than the estimated variation across officers. Additionally, the combined contribution of officer variables, θ_i and θ_{-i}, accounts for 2-5% of the total variation in arrest outcomes and 50-70% of the explained variation of the model, given a total model R^2 of 0.074 (Adjusted R^2 of 0.043) (See Table 2).

4.3 *Tests of Officer Sorting to Responses*

As discussed above, patrol officers responding to 9-1-1 calls have some scope to volunteer to respond to certain to incidents. While I control for a rich array of observable call characteristics, the estimates of individual officer arrest propensity could be biased if officers systematically respond to calls based on unobservable characteristics.

There are two potential patterns of selection. First officers with a high arrest propensity may be more *active*, both in the sense that they could respond to more calls or that they could be more likely to respond to marginal or less serious calls. First, this pattern would imply that high arrest officers should respond to more offenses than low arrest officers. Second, this pattern could create a negative correlation between officer effects, $\hat{\theta}_i$ and the error terms, ε_{ikgt}. This negative selection bias would deflate the dispersion in officer fixed effects and lead to an underestimate of this parameter.

To address this concern, I test whether officers with higher arrest propensities respond to more calls for service. The correlation between arrest propensity and the number of calls per officer is -0.1 (Table 2). This small negative relationship suggests that high arrest officers are not more *active* than low arrest officers. Moreover, the primary conclusion of this paper is that officers differ substantially in their arrest likelihood, so any potential negative selection related to high arrest officers responding to more marginal calls will lead to a more conservative estimate of the key parameter.

Alternatively, officers who have a high arrest propensity could prefer to respond to incidents with a higher likelihood of arrest, and officers who have a low arrest propensity could prefer to volunteer for incidents with a lower likelihood of arrest. In either case, the estimates of $\hat{\theta}_i$ will be positively correlated with the error terms ε_{ikgt} and inflate the dispersion in officer fixed effects.

First, I consider the importance of observable call characteristics that may affect officer response choices. Officers may choose to respond to calls based on call severity, dispatch code, availability of other officers working their shift and time within shift as well as the beat of the call, factors captured in X_{kt} and ϕ_g. In contrast, officers are assigned to shifts and match with vehicle partners, δ_{dt} and $\tilde{\theta}_{-i}$, before they can choose to respond to a particular call. In Figure 3.A, I calculate the correlation between the distribution of $\hat{\theta}_i$ from the full model to $\hat{\theta}_i'$ estimated from a model that omits call characteristics that could influence an officer's decision to respond to a call.[14] Perfect correlation between these estimates would imply that officer effects are orthogonal to this set of call characteristics in the model, or are nearly randomly assigned to calls. This figure shows that the $\hat{\theta}_i'$ is similarly distributed relative to the $\hat{\theta}_i$ estimates from the full model, with a standard deviation of 0.014 that is similar to the baseline standard deviation of 0.013. Additionally, the estimates across these distributions have a very high correlation of 0.93. This high correlation suggests that observable call choice characteristics are not very important controls in the estimation of the officer effects distribution. These tests are informative if unobservable offense characteristics are correlated with observables, an assumption that is often applied in tests of endogeneity.[15]

Next, I construct a balance test that relates officer effects to model estimates of predicted arrest likelihood for calls. Figure 3.B plots officer effects relative to average predicted

[14]The full model includes peer group effects for the full group of officers first dispatched to a call, potentially up to 4 officers from multiple cars. The peer group effect, $\tilde{\theta}_{-i}$, in the model without call decision characteristics is limited to any peer that is patrolling in the same car as the focal officer.

[15]Figure A3 plots the unconditional officer arrest propensity (centered) against the covariate-adjusted effects. It is notable that there is a very high correlation between these estimates of 0.85, even without including any pre-determined covariates in the comparison estimates.

arrest likelihood of call responses by officer, where arrest predictions are measured net of θ_i and θ_{-i}. There is no significant relationship between these variables, with a total correlation of 0.004.

As an additional test, I focus on two settings where officer sorting is less likely to impact the results, calls that occur when few officers are available, a "Low Availability" sample, and urgent calls, a "High Urgency" sample. In both settings, officers will be more constrained in their ability to volunteer to respond to a call. I define the "Low Availability" sub-sample by splitting the sample at the median officer availability rate within a general patrol shift cell (8 Hour Shift*Day of the Week*Division group), to account for variation in total staffing across shifts.[16] Similarly, I define "High Urgency" calls as those that have below median time between when a call is received and dispatched within a general patrol shift cell.

Figure 3.C and 3.D show the results of restricting the observations to the "Low Availability" and "High Urgency" sample. If dispersion in officer behavior is increased by officer sorting, we would expect the estimates of dispersion to be lower in these robustness samples than the baseline. However, the graphs show a strikingly close match between the distributions, with a correlation of approximately 0.93 between the officer effects in each of the robustness samples and their corresponding baselines. The estimated dispersion in the "Low Availability" and "High Urgency" samples is also similar to the baseline, with standard deviations of 0.013 and 0.014 respectively.[17] The evidence that officer dispersion is similar in samples where officer sorting is less likely reduces concerns that officer effects are systematically related to unobservable offense characteristics.

[16] I determine the total officers working on a shift and the rate of available officers using the full sample of all dispatched calls, including those that are excluded from the analysis given sample restrictions on number of responding officers, total officer responses or time to dispatch (as described in Section 2.2.

[17] I restrict each robustness sample to officers with at least 500 observations in the relevant sub-sample. I then compare estimates within these sub-samples to corresponding samples that include all observations for these officers. Appendix Table A1 shows that the characteristics of the robustness samples are similar to the primary analysis sample.

4.4 *Robustness Tests of the Empirical Model*

How large is the dispersion in officer arrest behavior? One way to assert that the distribution of officer arrest propensities is meaningful is to benchmark the observed standard deviation in officer effects, $\hat{\theta}_i$, to the amount of variation that would be observed across officers if there were no "true" officer effects. Even in the absence of officer differences, there will be some measured variation in outcomes across officers, simply due to idiosyncratic variation in the error term or "noise."

I use a bootstrap simulation to confirm that confirm that the results in this study reflect actual variation in behavior across officers. This test estimates a distribution of officer dispersion and the explanatory power of officer effects under the assumption that the "true" prediction value of all officer effects is 0 and reference the actual model estimate to this distribution.[18] Figure A4 displays the results of these tests using 100 bootstrap replications. The model estimate is well outside the 95% confidence interval given by the bootstrap tests. This graph confirms that the estimated variation in officer effects is not due to noise in the data.

Next, I test the robustness of the baseline model by considering several alternate specifications in Table A2. In column (2), I substitute the 234 police beat categories with narrower geographic area controls of 1,161 police reporting areas in the city of Dallas. Column (3) alternatively substitutes the police beat controls with geographic controls that vary by time, or 1,820 Police Sector*Month category variables. Column (4) includes 33,135 individual 8-hour shift level indicators (Date*8-hour shift*Division) instead of the monthly aggregated shift indicators, δ_{dt}, used in the primary specification. By conditioning on individual 8-hour shifts, this specification absorbs variation in arrests at the date by police

[18] In this test, I calculate the residuals, $\hat{r}$, and predicted outcome values, $\widehat{Arrest}$, from a regression that does not include officer fixed effects, thereby assuming a null hypothesis that the true value of all θ_i and θ_j terms is zero. I apply the wild cluster bootstrap (Cameron et al., 2008) to allow errors to be correlated within calls, permitting error correlation across duplicated call observations. I apply weights of $w \in \{-1, 1\}$ to residuals $\hat{r}$ that are constant within each call and construct a new outcome variable as the predicted outcome plus the weighted residual, $\widetilde{Arrest}_b = \widehat{Arrest} + w_b\hat{r}$. I then regress this new outcome variable, $\widetilde{Arrest}_b$, on the fully specified model that includes officer fixed effects and calculate the dispersion of the officer effects for each iteration. I have also conducted this test imposing the restriction that $\widetilde{Arrest}_b$ is binary in each iteration. To do this, I set the highest values of the bootstrap outcome variable equal to one such that the mean of $\widetilde{Arrest}_b$ equals the mean of $Arrest$. The results of this bootstrap test are similar and are available on request.

division geography level, accounting for factors such as changing weather conditions, holidays and other day specific events in the city of Dallas. In column (5), I replace the 22 aggregated dispatch codes used in the main specification with 85 specific dispatch code categories in the call data.[19] Across specifications these additional controls do little to change the analysis. In fact, the correlation between the base distribution of officer fixed effects and these alternate specifications is above 0.96.

Column (6) restricts the sample to the sub-set of dispatch calls with an offense report record, approximately 10% of calls. This smaller sample permits additional controls for complainant characteristics, including gender, race and whether there is a victim injury, as well as location type variables (e.g. Apartment Residence, Street, Business etc.), but has the precision cost of fewer observations per officer as well as issues of potential sample selection related to officer choices to designate incidents as offenses. The dispersion in this sample is comparable to the baseline model, with 1 standard deviation in officer effects corresponding to a 29% increase in arrest likelihood. Likewise, the total variation explained by officer effects is similar in this sample to the baseline, with 5.5% of total variation explained. The officer effects in this sample are positively correlated with the baseline estimates, with a correlation of 0.492.[20]

Columns (7) and (8) consider alternative procedures to adjust the estimates for precision instead of the Empirical Bayes' method used in the primary results. In column (7), I report the dispersion in unadjusted officer fixed effects from the first stage, where the sample is restricted to officers with more than 2,000 observations. In column (8), I weight the unadjusted officer fixed effects by the number of observations per officer. Across both of these alternative precision methods, the dispersion estimates are similar to the base model, with a 1 standard deviation in officer effects corresponding to a 33-36% increase in arrest probability. Lastly, in column (9), I report dispersion in officer behavior using the unadjusted first stage officer fixed effects. With no adjustment for precision, this standard deviation estimate is only slightly larger than the base model, corresponding to a 39% increase in

[19] I collapse dispatch code categories to increase the estimation speed of the model and address small cell categories in the majority of the analysis.

[20] Prior drafts of this project used the offense report sample for the primary analysis, given the rich controls in this data set. The current version focuses on the total sample of all dispatched calls to address potential selection in officer decisions to complete an incident offense report.

arrest probability.

4.5 *Officer Demographics*

How does officer arrest propensity relate to officer demographics? A natural next step is to consider how the estimated officer fixed effects, $\hat{\theta}_i$, are associated with officer demographic characteristics. Table 3 shows the results of regressing $\hat{\theta}_i$ terms on officer race, gender, age, trainee or sergeant status and experience.[21] These regressions offer information about whether officers with specific traits systematically differ in their arrest propensities.

Officers with more experience have lower arrest propensities. All else equal, the likelihood of arrest is 12% lower when a responding officer has 10 years of experience instead of 5 years of experience. Likewise, officers that are trainees are 6% more likely to make arrests than non-trainees. Conversely, the small share of sergeants in the sample have a higher arrest propensity, with 29% higher likelihood of arrests than non-sergeants.

Officer race also makes a difference in the likelihood of arrest; Black officers are 9% less likely to make arrests relative to White officers. Male and female officers do not have statistically different arrest propensities. The regressions are very similar using officer effects derived from the "Low Availability" and "Unlikely Response" samples.

Overall, demographic factors do not explain a very large share of the total variation in officer effects, with regression R^2 statistics of ≈ 0.1. This analysis shows that the majority of the variation in arrest behavior observed across officers is due to unobservable characteristics of police officers, which may include officer preferences and specific officer experiences.

5 Arrest Propensity and Characteristics of Arrests

Identifying meaningful variation in arrest outcomes across individual officers provides evidence about the existence and extent of police discretion in the field. However, arrests are not a normative outcome. An arrest may have positive or negative welfare consequences depending on the incident context, culpability of the arrestee, severity of the offense, impli-

[21] Salary is omitted from this regression because it is nearly perfectly correlated with experience, given the compensation formulas used by the department. Results are similar when salary is used instead of years of experience.

cations for public safety and the subsequent burden for the arrested individual and his/her family.

In this section, I examine relationships between officer arrest propensity and arrest characteristics, in order to better understand the margins of arrest decisions across officers. One possibility is that officers with high arrest rates may simply be more *productive* than low arrest officers. Alternatively, high arrest officers could have a lower severity or evidence threshold for making an arrest and/or be more *aggressive* than low arrest officers. The analysis in this paper cannot explicitly determine the optimal level of arrests or arrest characteristics from a welfare perspective. Instead, I analyze arrest characteristic measures as relative indicators of the severity of arrest offenses and sanctions as well as adverse consequences of arrest events, including use-of-force.

Specifically, I draw on multiple data sources from DPD and Dallas County on arrests, offense reports, jail stays, county court outcomes and non-shooting use-of-force incidents to construct measures of arrest characteristics for each officer. I estimate arrest characteristics at the officer-level for two reasons. First, some characteristics such as whether an arrest results in use-of-force are too rare (less than 2% of arrests) to effectively use as an outcome in the sample of 9-1-1 calls. Second, the multiple data sources covering arrest characteristics have differing limitations, levels of detail and sample coverage (see description in Table 4 and Online Data Appendix A5 for additional details on data features and cleaning).

I examine associations between officer arrest propensity and arrest characteristics by regressing these various officer-level arrest characteristic measures on estimates of officer arrest effects, $\hat{\theta}_i$. I use estimates of $\hat{\theta}_i$ from a two-year training sample period and outcomes from a non-overlapping test sample period in order to address concerns about the joint determination of these variables.[22] As discussed above, $\hat{\theta}_i$ can be viewed as estimates of each officer's permanent underlying arrest propensity, as they are derived from a sample of interactions that are initiated by civilians and not officers, and adjusted for incident characteristics, geographic and time factors and peer influence. These features make $\hat{\theta}_i$ a

[22]The training sample period for this analysis is the first and last year of the sample, or 7/2014-6/2015 and 11/2017-10/2018. Officer effects are estimated using the fully specified model described in Section 3 for officers with more than 500 observations in the training sample period. The test sample period is 7/2015-10/2017. I chose this bookended training data set to maximize the coverage of the data sets available for arrest characteristics, most of which end in the beginning of 2018.

valid covariate in measuring associations between arrest propensity and arrest characteristics.

As a first step, I check whether the training sample $\hat{\theta}_i$ are reasonably associated with baseline characteristics of total officer arrest rates and crime offenses in Panel A of Table 5. The first row shows that $\hat{\theta}_i$ measured in the training sample is strongly positively correlated to a similarly defined officer effect in the test sample, with a correlation of 0.6. Similarly, the training sample $\hat{\theta}_i$ is strongly associated with unconditional arrest rates and total arrests in the test sample (Rows (2) and (3)), suggesting a high level of persistence in officer arrest propensity. As noted in the officer sorting analysis in Section 4.3, officer arrest propensity is not substantively related to the number of calls an officer responds to, with a slight negative relationship (Row (4)). Lastly, I test whether officer arrest propensity is related to crime exposure in Rows (5) and (6). I do this by constructing measures of crime exposure for each officer as weighted averages of monthly violent and property crimes reported to police relative to population (per 100,000 residents) in the geographic police division-months where an officer has worked.[23] Officer arrest propensity is unrelated to violent crime exposure, and officer arrest propensity is slightly positively correlated with property crime exposure, though this effect is very small in relative terms. These tests support the finding that high and low arrest officers do not differ in their arrest outcomes because they face offenses that differ in severity.

Next, I examine relationships between officer arrest propensity and characteristics of arrest outcomes in Panel B of Table 5. Row (7) shows that high arrest officers are significantly more likely to use physical force during the course of an arrest, where use-of-force outcomes are scaled by total arrests for each officer. I construct a conservative definition of a use-of-force incident as an arrest where an officer used physical force or the civilian was injured, the civilian did not resist and was not armed, and the officer was not injured. A 1 standard deviation increase in officer arrest propensity is associated with a 20% increase in officer use-of-force arrest rates.

Rows (8) and (9) show that high arrest officers make more misdemeanor arrests than low arrest officers, but have similar felony arrest rates. This pattern implies that mis-

[23]Violent crimes include aggravated assault, other assault, robbery, murder/manslaughter and weapons charges. Property crimes include burglary, theft, vehicle theft, embezzlement, fraud, forgery and counterfeiting. Population data at the census tract level is used to calculate division-level population.

demeanor arrests involve a greater level of officer discretion and subjectivity than felony arrests.[24]

Despite having a larger share of lower severity arrests, high arrest officers are also more likely to book arrestees in county jail after completing an arrest. Officers have some agency over whether arrested individuals are booked in the county jail, particularly for lower level offenses.[25] Row (10) shows that given the option, high arrest officers are more likely to impose a higher sanction on arrested individuals. The estimates imply that a 1 standard deviation increase in officer arrest propensity corresponds to a 10% increase in the proportion of arrests that result in a jail stay.

In Texas, county jail bookings are intended for offenses that are more serious than Class C Misdemeanors, which are legally punishable with a fine of up to $500 and no jail time (Texas, 1993). However, in practice, officers can exercise their judgment in booking individuals with Class C Misdemeanors in jail, and may do so if they anticipate that a higher level charge will be assigned and upheld for the arrestee after booking. In cases when a higher level charge is not upheld, these individuals will be released from jail because their low-level offenses do not legally warrant jail time. Row (11) measures the proportion of an officer's jail admissions that resulted in release because an individual was booked for a Class C Misdemeanor and a more serious charge was not upheld. The regression coefficient shows that high arrest officers are more likely to book arrestees in county jail for very low level offenses that legally do not warrant jail time. This result is consistent with the result in Row (10) and shows that high arrest officers impose harsher sanctions on arrestees than low arrest officers.

Lastly, I consider how officer arrest propensity relates to arrest dismissal and conviction rates in Rows (12) and (13).[26] High arrest officers do not significantly differ from low

[24]These tests display the proportion of total arrests that are either felonies or misdemeanors, where these categories are not mutually exclusive (an incident may result in both a felony and misdemeanor arrest) and 46% of arrests are not coded as either felonies or misdemeanors.

[25]Arrests for minor offenses may alternatively result in detention in the Dallas City Marshal Detention Center or a citation with promissory notice to appear in court. Additionally, some arrests may result in transfer of an individual to another police agency or mental health facility.

[26]This rate are restricted to an arrestee sample that includes information on charge severity level and excludes Class C Misdemeanors. The rate is further restricted to the sample of arrestee records that can be linked to court records using a fuzzy match by name and offense date ($\approx$ 90% match rate within charge severity restricted sample).

arrest officers in terms of ultimate court dismissal or conviction outcomes for arrests. This lack of a relationship could be related to the fact that dismissals and negotiated conviction plea deals are nearly always determined by prosecutor discretionary decisions which may imperfectly measure charge quality.

This analysis provides evidence that high arrest officers are more likely to make arrests for lower level offenses, are more likely to apply harsher punishments and are more likely to use force during an arrest than low arrest officers. Collectively, this pattern of arrest characteristics suggests that the average quality of an arrest is likely to be lower for high arrest vs. low arrest officers.

6 Racial Bias Among Officers

6.1 Racial Disparity in Arrest Outcomes

Racial disparity and racial bias are critical dimensions of evaluating police interactions. A multitude of research on policing and the criminal justice system shows that the Black, Hispanic and low-income individuals comprise the majority of arrests and that the arrest shares for these groups exceed their proportion of the population. Dallas is no exception to this rule: Black arrestees 50% of total offense arrests and only 24% of the Dallas population, adjusted for the proportion of arrests with demographic information (Census, 2015).

It is important to note that across the distribution of officer arrest propensity, all officers have a vastly higher proportion of Black arrests than White or Hispanic arrests (Figure 4.A). Similarly, each officer race group has a higher proportion of Black arrests than arrests for other civilian races (Figure 4.B). This pattern could be consistent with statistical discrimination, institutional discrimination or uniform taste-based bias that is common across all officers. Institutional racial discrimination will occur when the organizational priorities of the department direct resources toward policing one race group relative to others, and all officers behave similarly given these priorities. Statistical discrimination will occur when civilians of a particular race have a higher propensity to engage in criminal behavior, and race is used by officers as a signal of criminal activity. At the same time, the higher representation of Black arrestees could also be consistent with uniform officer attitudes of taste-based racial bias against Black civilians. Recent evidence on the importance of implicit racial bias in decision-making could be consistent with uniform taste-based discrimination against minority groups (e.g. Eberhardt et al., 2004).

6.2 *Testing for Taste-Based Racial Bias*

In this section, I apply a test of taste-based racial bias by leveraging variation in race-specific arrest outcomes across officers. I adapt a test of racial bias used in Anwar and Fang (2006) examining officer bias in traffic stops across officer race groups. Their test finds evidence of racial bias when the relative ranking of officer arrest rates across officer groups changes within different suspect race groups. For example, their test finds evidence of bias if White officers have higher arrest rates than Black officers for Black suspects *and* Black officers have higher arrest rates than White officers for White suspects. In this case, either Black officers or White officers are racially biased (or both).

Their model allows officers of different groups to behave differently from one another, or have *non-monolithic* costs of effort, as long as these differences are independent of suspect race. For example, if one officer group is more likely than another officer group to arrest a suspect of any race, this does not necessarily imply that either group is racially biased. This paper provides strong evidence that officers have *non-monolithic* costs of making arrests, as there is significant variation in total arrest propensity across individual officers.

Critically, the test also does not find evidence of racial bias if *all officers* arrest individuals in some suspect race groups more than others. This feature allows officers to statistically discriminate against suspects, by using suspect race as a signal of offending characteristics that are correlated with race. Arrest rates can differ across suspect race groups in the test because differences in officer group arrest rates are always measured as a relative ranking within a suspect race group.

I consider officer group categories as either arrest propensity types, defined as deciles of the officer effects or $\hat{\theta}_i$ distribution, or officer race groups, as is typical in this literature. I then consider whether the ranking of race-specific arrest rates across officer groups is the same for all arrestee race groups.

In the setting of calls for service, not all call responses will necessarily have a potential suspect. I adapt the test to consider unconditional arrestee race outcomes relative to total calls, defined as 1 if a call resulted in an arrest of a person of particular race and 0 otherwise. In (Appendix A4), I show that the logic of the relative rank order test used in Anwar and Fang (2006) applies to these unconditional arrestee race outcomes.

Specifically, I conduct the racial bias test by regressing unconditional race-specific arrest outcomes (e.g. arrest occurred and the arrestee was Black) on the full empirical model described in Section 3. From this model I recover estimates $\hat{\theta}_{i,r}$ or arrest propensity officer effects for a particular arrestee race outcome. These officer effects are estimated net of call characteristics, geography, timing and peer influence. I then regress the $\hat{\theta}_{i,r}$ estimates for each arrestee race on officer group category indicators, allowing comparisons of whether the rank order across officer groups is constant for different arrestee races.

A major advantage of this approach is that I am able to test for racial bias among officers in a setting that is not affected by officers electing to initiate interactions. With the exception of West (2015), which studies racial bias of state troopers who are randomly dispatched to motor vehicle accidents, prior work studying racial bias in policing has typically examined interactions between officer and suspect race in officer-initiated incidents, such as traffic stops (e.g. Horrace and Rohlin, 2016; Anbarci and Lee, 2014; Antonovics and Knight, 2009; Anwar and Fang, 2006; Grogger and Ridgeway, 2006). These papers consider suspect race as a given characteristic of a traffic stop; however, in reality, suspect race is also a choice variable of the officer, who chooses to stop a particular individual.

An additional strength of the test is the fact that the estimates of arrest rates for each officer group and arrestee race are adjusted for contextual factors through the regression in the first stage. This mitigates concerns that particular officer groups are systematically related to incident characteristics in a way that would differentially expose officer groups to different types of potential suspects. This challenge is non-trivial in the case of the officer race groups used in Anwar and Fang (2006) who employ a re-sampling procedure to balance the geographic areas for officers of different races. In the setting of Dallas, officers are more likely to work in divisions with a high proportion of residents of their own race, which complicates an unadjusted officer race group comparison. The regression framework improves on geographic re-sampling by adjusting race-specific arrest rates for multiple dimensions of incident context in addition to geography, including timing, call severity and urgency, call type and peer presence.

Lastly, I am able to use officer identifiers to trace the distribution of officer effects by officer race (arrest propensity) group for each of the arrestee race outcomes. Plotting this distribution reveals the relative importance of behavioral differences within officer race (arrest propensity) group versus across officer race (arrest propensity) groups.

The racial bias test used in this paper can only detect taste-based racial bias to the extent that this bias differs across officer groups, a limitation of prior tests as well. The test is also unable to separate uniform taste-based racial bias from institutional racial bias or statistical discrimination.

6.3 *Results of the Test for Taste-Based Racial Bias*

The results of the racial bias test are striking. Table 5 shows the results of the test using officer arrest propensity groups (Panel A) and officer race groups (Panel B). The results show a strong and consistent ordering of officer groups across arrestee race. In Panel A, officers with higher arrest propensities are more likely to arrest individuals of any race group relative to officers with lower arrest propensities. In Panel B, Black officers are less likely to arrest individuals of any race group and there are no significant differences between Hispanic and White officers (omitted category). Given this near perfect mapping in relative officer group ranking for the different arrestee race groups, the test cannot reject the null hypothesis of no taste-based racial bias among officers. These results are robust to including a full set of available officer demographic controls in Table A4.

Figures A5 and A6 plot the officer effects for each arrestee race outcome, across officer arrest propensity groups and officer race groups. In Figure A5, the officer arrest propensity groups clearly separate the total officer distribution, with a consistent ordering across outcomes. In contrast, Figure A6 shows that officer race groups virtually overlap for each arrestee race outcome. The variation in arrestee outcomes across officer race groups is minuscule when compared to variation *within* officer race. The estimates imply that artificially fixing all officers to have the same race (White), would reduce the dispersion in the officer effect distribution by less than 1%.

These results could be consistent with reforms adopted by the Dallas Police Department prior to and during the sample period, including implicit racial bias training, de-escalation training, the use of body-worn cameras and sharing data on its operations with the public.

The racial bias test used in this paper has the advantage of being unlikely to yield a false positive claim of bias, an attractive feature given that a finding of racial bias can be a political flash point. However, as discussed above, the test used in this paper cannot detect taste-based racial bias in cases when all officers exhibit similar bias toward a particular group. In this setting, there are large race disparities in arrest outcomes for all officers, which could be consistent with uniform taste-based bias, implicit bias, institutional bias and/or statistical discrimination.

7 Conclusion

Individual police officers are critical to the outcomes of police work. This paper finds substantial variation in arrest behavior across officers responding to civilian-initiated calls for service, even after controlling for detailed characteristics of call incidents.

Analyzing high frequency data on calls for service in Dallas, Texas, I find that a 1 standard deviation increase in officer arrest propensity increases the likelihood of an arrest by 37%. The dispersion in predicted arrest outcomes across officers is similar to the dispersion across geography but is approximately a third of the dispersion attributable to call type. While high and low arrest officers face similar crime offending environments, high arrest officers are more likely to use physical force during an arrest, arrest individuals for lower level offenses and impose harsher sanctions for low level offenses. These results suggest that high arrest officers may have a lower severity or evidence threshold for making arrests and applying sanctions.

I also adapt a relative test of racial bias leveraging differences in race-specific arrest outcomes across officers with different arrest propensities. Despite documenting large variation in total arrest behavior across officers, I fail to find conclusive evidence of taste-based racial bias in this setting. Instead, I find patterns consistent with officers having different net benefits of making arrests of civilians of any race; high arrest officers are more likely to arrest individuals of any race group than low arrest officers, and Black officers are less likely to arrest individuals of any race group than Hispanic or White officers.

Having established that individual officers are critical to the outcomes of police interactions, questions remain for future research. First, this project provides new evidence that police discretion is related to the characteristics of arrest outcomes. Future research should extend these findings to quantify the welfare costs and benefits of different types of arrests. A greater understanding of these welfare costs would allow policymakers to articulate police priorities and weigh the merits of alternative policing protocols.

Second, investments in reducing dispersion in officer behavior could yield benefits in the form of increased trust in law enforcement and equal access to police protection services. Future work should also assess the costs and benefits of different law enforcement practices that may be used to increase uniformity in officer behavior, including additional police training, monitoring procedures, mentorship programs and targeted hiring and firing of officers.

References

Aaronson, D., L. Barrow, and W. Sander (2007). Teachers and student achievement in the Chicago Public High Schools. *Journal of Labor Economics 25*(1), 95–135.

Anbarci, N. and J. Lee (2014). Detecting racial bias in speed discounting: Evidence from speeding tickets in Boston. *International Review of Law and Economics 38*, 11–24.

Antonovics, K. and B. G. Knight (2009). A new look at racial profiling: Evidence from the Boston Police Department. *The Review of Economics and Statistics 91*(1), 163–177.

Anwar, S. and H. Fang (2006). An alternative test of racial prejudice in motor vehicle searches: Theory and evidence. *American Economic Review 96*(1), 127–151.

BlueLivesMatter (2016, July). Opinion: Why do cops always support other officers? https://bluelivesmatter.blue/.

Brayne, S. (2017). Big data surveillance: The case of policing. *American Sociological Review (Forthcoming)*.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2008). Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics 90*(3), 414–427.

Census (2015). American community survey: Demographic and housing estimates. Technical report, Census Bureau, United States.

Chetty, R., J. N. Friedman, and J. E. Rockoff (2014). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review 104*(9), 2593–2632.

Correia, S. (2016). A feasible estimator for linear models with multi-way fixed effects. *Working Paper*.

Department of Justice (2017, March). Supporting Federal, State, Local and Tribal Law Enforcement. Memorandum, Office of the Attorney General, Department of Justice, United States.

Eberhardt, J., P. Goff, V. Purdie, and P. Davies (2004). Seeing black: Race, crime, and visual processing. *Journal of Personality and Social Psychology 87*(6), 876–893.

FBI (2015). Crime in the United States: Offenses Known to Law Enforcement. Technical report, Federal Bureau of Investigation, Department of Justice, United States.

Frydl, K. and W. Skogan (2004). *Fairness and Effectiveness in Policing: The Evidence*. National Academies Press.

Fryer, R. (2016). An empirical analysis of racial differences in police use of force. *National Bureau of Economic Research* (22399).

Gelman, A., J. Fagan, and A. Kiss (2007). An analysis of the New York City police department's "stop-and-frisk" policy in the context of claims of racial bias. *Journal of the American Statistical Association 102*(479), 813–823.

Goncalves, F. and S. Mello (2017). Racial bias in policing: Evidence from speeding tickets. *Working Paper*.

Grogger, J. and G. Ridgeway (2006). Testing for racial profiling in traffic stops from behind a veil of darkness. *Journal of the American Statistical Association 101*(475), 878–887.

Guarino, C., M. Maxfield, M. Reckase, P. Thompson, and J. Wooldridge (2015). An evaluation of Empirical Bayes' estimation of value-added teacher performance measures. *Journal of Educational and Behavioral Statistics 40*(2), 190–222.

Haugh, C. (2016, July 9). How the Dallas Police Department Reformed Itself. *The Atlantic*.

Horrace, W. C. and S. M. Rohlin (2016). How dark is dark? bright lights, big city, racial profiling. *The Review of Economics and Statistics 98*(2), 226–232.

Joh, E. (2016). Surveillance discretion: Automated suspicion, big data, and policing. *Harvard Law and Policy Review 10*, 15–42.

Jones, J. (2015, June). In U.S., confidence in police lowest in 22 years. Technical report, Gallup.

Kane, T. and D. Staiger (2008). Estimating teacher impacts on student achievement: An experimental evaluation. *National Bureau of Economic Research* (14607).

Knowles, J., N. Persico, and P. Todd (2001). Racial bias in motor vehicle searches: Theory and evidence. *Journal of Political Economy 109*(1), 203–229.

Koedel, C., K. Mihaly, and J. E. Rockoff (2015). Value-added modeling: A review. *Economics of Education Review 47*, 180–195.

La Vigne, N., J. Fontaine, and A. Dwivedi (2017, February). How do people in high-crime, low-income communities view the police? Technical report, Urban Institute.

Mas, A. and E. Moretti (2009). Peers at work. *The American Economic Review 99*(1), 112–145.

Mastrofski, S. (2004). Controlling street-level police discretion. *The Annals of the American Academy of Political and Social Science 593*, 100–118.

Morris, C. (1983). Parametric Empirical Bayes inference: Theory and applications. *Journal of the American Statistical Association 78*(381), 47–55.

Nickels, E. (2007). A note on the status of discretion in police research. *Journal of Criminal Justice 35*, 570–578.

Schneider, B. (2015, January). Do Americans trust their cops to be fair and just? New poll contains surprises. *Reuters*.

Silver, D. (2016). Haste or waste? Peer pressure and the distribution of marginal returns to health care. *Working Paper*.

Texas (1993). Penal code. *73rd Texas Legislature* (Title 3, Chapter 12.23).

Tsiaperas, T. (2016, June). Top Dallas cops take classes to learn how to fight bias. *The Dallas Morning News*.

West, J. (2015). Racial bias in police investigations. *Working Paper*.

Tables and Figures

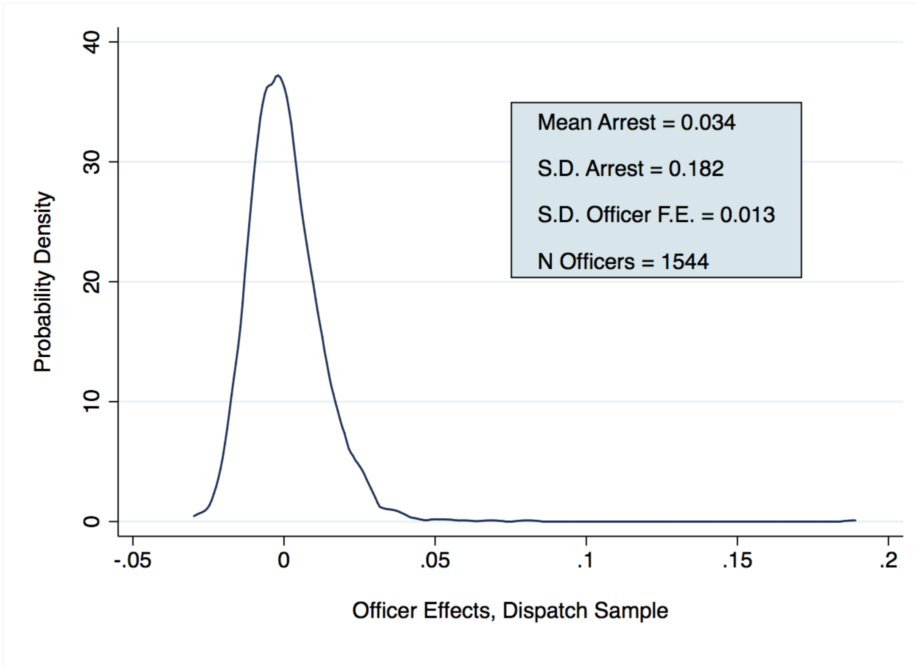
Table 1: Summary Statistics

	Total Sample		***Analysis Sample***	
	Mean	S.D.	Mean	S.D.
Total Observations	4,550,987		3,425,027	
Total Call Responses	2,444,520		1,987,467	
Total Officers	3,851		1,544	
Primary Outcome: Arrest	0.037	(0.189)	0.036	(0.187)
Black Arrestee	0.017	(0.129)	0.016	(0.126)
Hispanic Arrestee	0.007	(0.086)	0.008	(0.086)
White Arrestee	0.009	(0.092)	0.008	(0.091)
Call Characteristics				
Time to Dispatch (Minutes)	31.72	(99.52)	14.56	(19.65)
Proportion of Officers Unavailable	0.282	(0.153)	0.280	(0.153)
Number of Responding Officers	2.19	(1.05)	1.97	(0.682)
Call Priority				
High Priority	0.600	(0.490)	0.647	(0.478)
Low Priority	0.400	(0.490)	0.353	(0.478)
Dispatch Code Type				
Criminal Assault	0.008	(0.089)	0.003	(0.059)
Armed Encounter/Active Shooter	0.017	(0.130)	0.013	(0.114)
Robbery	0.018	(0.134)	0.018	(0.133)
Burglary of Business	0.042	(0.200)	0.040	(0.196)
Burglary of Vehicle	0.022	(0.147)	0.020	(0.141)
Burglary of Residence	0.058	(0.234)	0.056	(0.230)
Unauthorized Use of Vehicle	0.012	(0.110)	0.010	(0.097)
Theft	0.016	(0.125)	0.012	(0.110)
Criminal Mischief	0.111	(0.314)	0.114	(0.318)
Major Disturbance	0.271	(0.445)	0.299	(0.458)
Injured Person	0.006	(0.080)	0.005	(0.074)
Accident	0.117	(0.322)	0.119	(0.323)
Other	0.300	(0.458)	0.290	(0.454)
Time of Day				
Day (8am-4pm)	0.315	(0.464)	0.335	(0.472)
Evening (4pm-12am)	0.427	(0.495)	0.423	(0.494)
Overnight (12am-8am)	0.258	(0.438)	0.242	(0.428)
Officer Characteristics				
Total Calls	2,593.9	(1,261.5)	2,548.7	(947.1)
Trainee	0.080	(0.272)	0.067	(0.250)
Sergeant	0.036	(0.186)	0.009	(0.095)
Salary ($10,000s)	57.94	(11.36)	57.30	(10.60)
Years of Experience	11.86	(8.62)	11.66	(8.48)
Age	37.31	(9.77)	37.23	(9.70)
Female	0.150	(0.357)	0.142	(0.349)
Black	0.255	(0.436)	0.260	(0.439)
Hispanic	0.217	(0.412)	0.213	(0.410)
White	0.486	(0.500)	0.484	(0.500)

This table displays summary statistics of the data used in analysis. The first column, "Total Sample", consists of all offenses reported through calls for service in the data, with each offense incident duplicated for each initial police officer responder. The second column, "Analysis Sample", summarizes the primary analysis sample and excludes records for police officers that respond to fewer than 1,000 calls, calls dispatched over 90 minutes after the call was received and call responses with more than 4 initial responders.

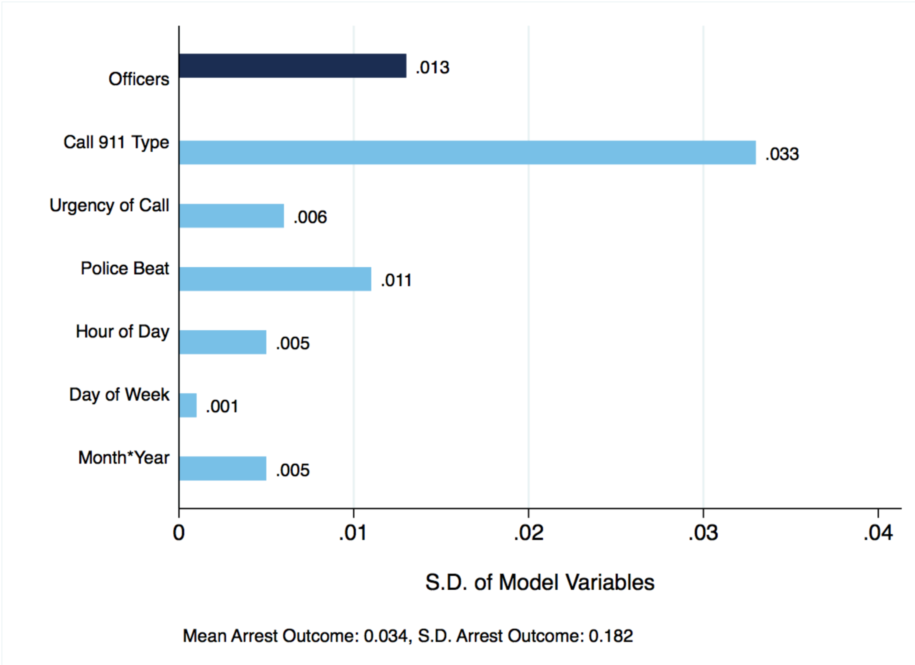
Figure 1: Dispersion in Officer Arrest Propensity

Figure 1.A: Distribution of Officer Effects, $\hat{\theta}_i$



The top figure graphs the distribution of the estimated Officer Effects, $\hat{\theta}_i$, measured using the arrest outcome model on the analysis sample. Each officer in the sample has at least 1,000 call responses.

Figure 2: Distribution of Officer Effects Relative to Other Model Components



This figure plots the distribution of the Officer Effects, $\hat{\theta}_i$, relative to the distribution of other estimated model coefficient groups. These dispersion estimates are derived from simplified variant of the baseline model in order to easily separate time vs. geographic coefficients. The alternate model version does not affect the estimate of the standard deviation of Officer Effects. This variant is $Arrest_{ikgt} = \theta_i + \theta_{-i} + \pi X_{kt} + D_{1,month*year} + D_{2,day-of-week} + \phi_g + \varepsilon_{ikgt}$, where δ_{dt} Shift Fixed Effects (8 Hour Shift*Day of the Week*Division*Month*Year) are substituted with more general component fixed effects for hour of day (in X_{kt}), Month*Year Fixed Effects, Day of the Week Fixed Effects and Beat Fixed Effects (ϕ_g) which are collinear sub-units of police divisions.

Figure 3: Tests of the Importance of Officer Sorting to Officer Effect Distribution

Figure 3.A: Officer Effects Measured with and without Officer Decision Controls

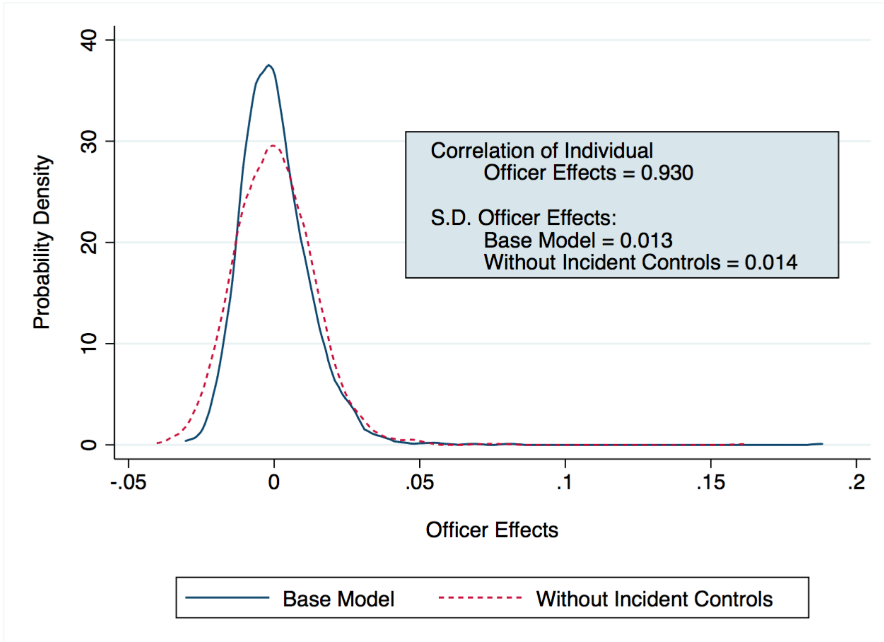
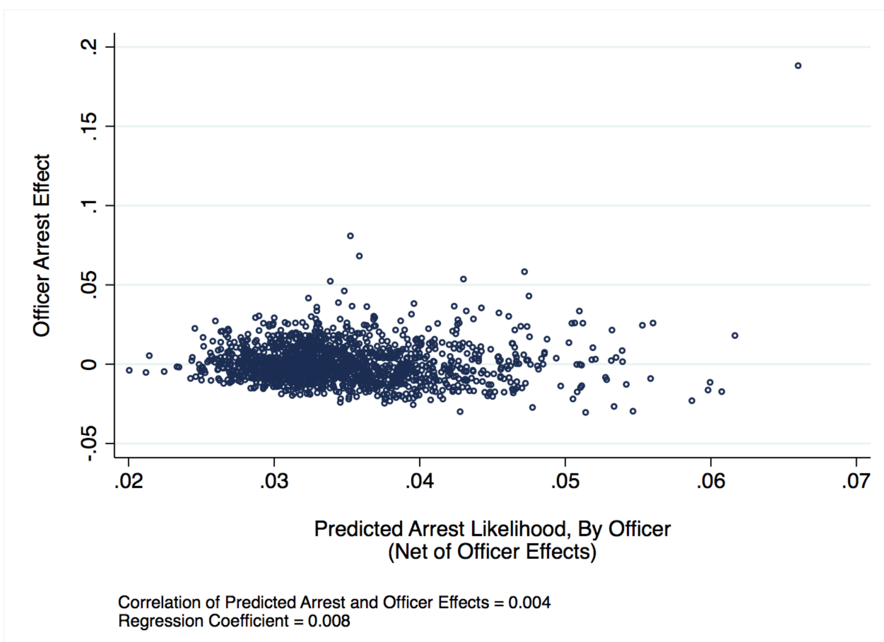


Figure 3.B: Officer Effects and Predicted Arrest Score



The top figure compares the base model officer effects, $\hat{\theta}_i$, to officer effects, $\hat{\theta}'_i$, that are estimated from a model that does not include call characteristics and police beat fixed effects, X_{kt} and ϕ_g. Additionally, θ_{-i} controls for other officers responding to a call are replaced with θ'_{-i} or controls for an officer that is patrolling in the same car as the focal officer. The omitted call characteristics in this figure represent margins that an officer may choose at the level of a call response, as shifts and same-car partners are determined prior to a call event. The bottom figure plots the estimated Officer Effects, $\hat{\theta}_i$, against the average predicted arrest score of each officer's calls. Predicted arrest scores are calculated as the model prediction of arrest for each call, net of responding officer and peer officer effects, θ_i and θ_{-i}. These predicted scores are averaged across responses for each officer in the sample.

Figure 3: Tests of the Importance of Officer Sorting to Officer Effect Distribution (Continued)

Figure 3.C: Officer Effects in Full Sample and Low Availability Sample

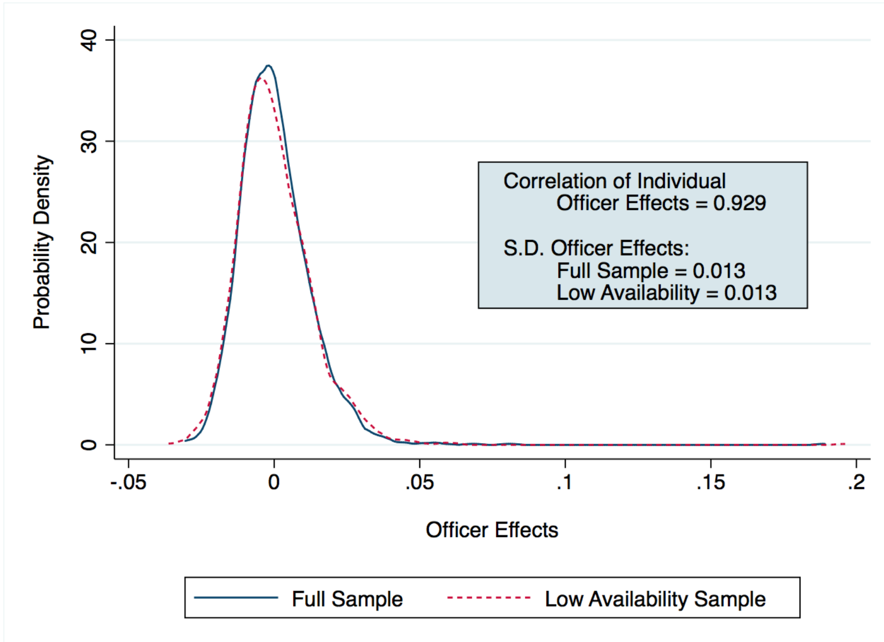
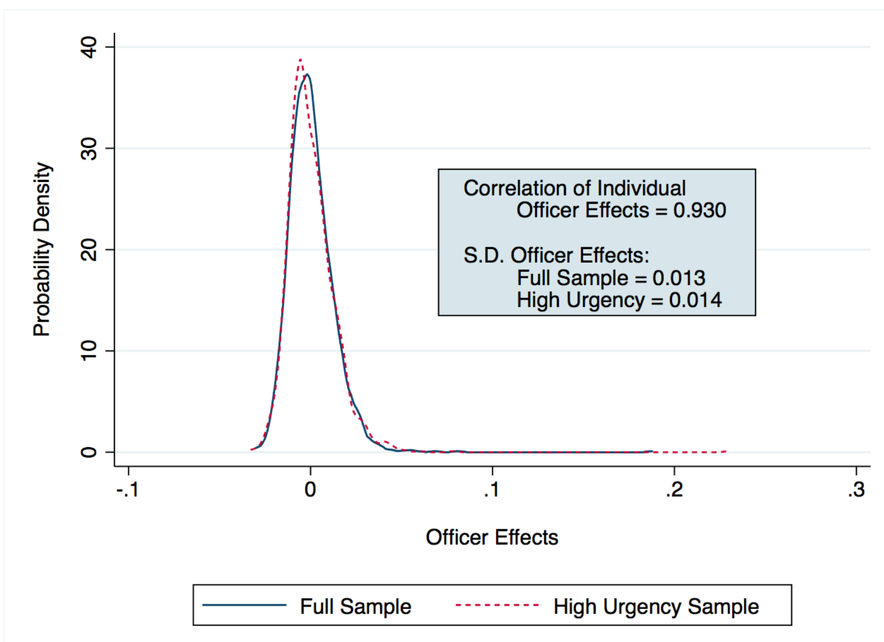


Figure 3.D: Officer Effects in Full Sample and High Urgency Sample



The "Low Availability" sub-sample is determined by taking the set of observations where a greater proportion of officers are unavailable because they are responding to other offenses at the time an offense is dispatched, split at the median within broad patrol shift cells (8 Hour Shift*Day of the Week*Division Group). The "High Urgency" sub-sample consists of observations with a shorter time between when a call is received and when it is dispatched, split at the median within broad patrol shift cells (8 Hour Shift*Day of the Week*Division Group). The analysis is restricted to officers with at least 500 observations in the sub-samples. The corresponding base sample benchmark is estimated over the full set of responses for the same officer group.

Table 2: Tests of the Importance of Officer Sorting to Officer Effect Distribution

	(1) Full Sample	(2) Full Sample, Without Incident Controls	(3) Low Availibility Responses	(4) High Urgency Responses
Total Officers	1,544	1,544	1,509	1,542
Total Observations	3,206,609	3,421,379	1,551,377	1,807,302
Mean of Outcome	0.034	0.036	0.033	0.034
S.D. of Outcome	0.182	0.187	0.179	0.180
Distribution of Officer Effects				
S.D. of Officer Effect	0.013	0.014	0.013	0.014
% Change: 1 S.D. Increase in Officer Effect	37.2%	39.4%	40.3%	40.4%
Gap: 10th to 90th Percentile in Officer Effect	0.029	0.034	0.030	0.029
% Change: 10th to 90th Percentile in Officer Effect	83.9%	93.7%	89.4%	86.5%
Contribution of Officer Effects				
Total R-2	0.074	0.018	0.080	0.089
Total Adjusted R-2	0.043	0.015	0.045	0.047
Additional R-2 from Officer Effects	0.051	0.006	0.054	0.062
Additional Adjusted R-2 from Officer Effects	0.023	0.004	0.023	0.024
Relative % of R-2 from Officer Effects	69.4%	30.6%	66.7%	69.9%
Relative % of Adj. R-2 from Officer Effects	53.1%	29.6%	50.1%	51.5%
Correlation of Officer Effects				
Full Samples		0.930	0.929	0.930
Correlation of Officer Effects & Number of Responses	-0.100	-0.111	-0.085	-0.108
Correlation of Officer Effects & Predicted Arrest Score	0.004	-0.038	0.005	0.070

This table summarizes the main analysis arrest results and tests of officer sorting. The "Without Incident Controls" model is estimating excluding call characteristics and police beat fixed effects, X_{kt} and ϕ_g. Additionally, θ_{-i} controls for other officers responding to a call are replaced with θ'_{-i} or controls for an officer that is patrolling in the same car as the focal officer. The "Low Availability" sub-sample is determined by taking the set of observations where a greater proportion of officers are unavailable because they are responding to other offenses at the time an offense is dispatched, split at the median within broad patrol shift cells (8 Hour Shift*Day of the Week*Division Group). I determine the "High Urgency" sub-sample by taking the set of observations with a shorter time between when a call is received and when it is dispatched, split at the median within broad patrol shift cells (8 Hour Shift*Day of the Week*Division Group). The "Contribution of Officer Effects" measures the additional R^2 from adding θ_i and θ_{-i} officer controls to a model with all other baseline controls. The "Correlation of Officer Effects" for "Full Samples" is the correlation between the alternatively estimated Officer Effects for officers across columns. The "Correlation of Officer Effects & Predicted Arrest Score" is the relation between estimated Officer Effects and the average predicted arrest score for each officer (across call responses) estimated from the model, net of officer controls θ_i and θ_{-i}. The observation count across the samples excludes singletons dropped from the model given the large number of controls and fixed effects and are therefore lower than the observation counts in the summary statistics.

Table 3: Officer Effects and Officer Demographics

Outcome: Officer Effect	***Analysis Sample*** (1)	***Low Availability Sample*** (2)	***High Urgency Sample*** (3)
Black	-0.003***	-0.003***	-0.003***
	(0.001)	(0.001)	(0.001)
Hispanic	-0.001	-0.001	0.000
	(0.001)	(0.001)	(0.001)
Other Race	-0.003*	-0.002+	-0.002
	(0.001)	(0.001)	(0.001)
Female	0.000	0.000	0.001
	(0.0007)	(0.0008)	(0.0008)
Sergeant	0.010***	0.008**	0.008*
	(0.00276)	(0.00279)	(0.00313)
Trainee	0.002*	0.001	0.003**
	(0.001)	(0.001)	(0.001)
Age	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)
Experience	-0.001***	-0.001***	-0.001***
	(0.0002)	(0.0002)	(0.0002)
Experience^2	0.00002***	0.00002***	0.00002***
	(0.00000)	(0.00001)	(0.00001)
Observations	1,542	1,507	1,540
R-squared	0.117	0.106	0.100
Adjusted R-squared	0.112	0.101	0.095
Fixed Effect Mean	0.000	0.000	0.000
Fixed Effect S.D.	0.013	0.013	0.014
Outcome Mean	0.034	0.033	0.034
Outcome S.D.	0.182	0.179	0.180

*** $p<0.001$, ** $p<0.01$, * $p<0.05$, + $p<0.1$
This table shows regression results of Officer Effects, $\hat{\theta}_i$, regressed on fixed officer characteristics, at the officer level. Robust standard errors are in parentheses. The analysis in columns (2) and (3) is restricted to officers with at least 500 observations in each sub-sample. Other race officers are the omitted race category. Officers without demographic information are excluded from the regressions.

Table 4: Officer Effects and Arrest Characteristics

Outcomes	Observations	Mean	Coefficient	Standard Error	*Change 1 S.D. Officer Effect*
A. Baseline Characteristics					
(1) Officer Effect, Test Sample	1,257	0.000	0.614***	(0.029)	21.4%
(2) Unconditional Arrest Rate	1,488	0.039	0.862***	(0.041)	27.8%
(3) Total Arrests	1,488	49.99	988.0***	(72.50)	24.9%
(4) Total Observations	1,488	1,354	-2,534+	(1,359)	-2.4%
(5) Violent Crime (100,000 Residents)	1,488	81.83	40.65	(36.26)	0.6%
(6) Property Crime (100,000 Residents)	1,488	298.5	299.1*	(146.2)	1.3%
B. Arrest Characteristics					
(7) Use of Force Arrests (Total Arrests)	1,414	0.017	0.276***	(0.078)	20.6%
(8) Felony Arrests (Total Arrests)	1,529	0.149	-0.147	(0.153)	-0.2%
(9) Misdemeanor Arrests (Total Arrests)	1,529	0.407	2.259***	(0.355)	7.0%
(10) Jail Bookings (Total Arrests)	1,463	0.360	2.496***	(0.402)	8.7%
(11) Class C Misdemeanor Jail Releases (Jail Bookings)	1,465	0.150	0.866***	(0.261)	7.2%
(12) Dismissed/Non-conviction Arrests (Matched Arrests)	1,488	0.170	0.160	(0.286)	1.2%
(13) Conviction Arrests (Matched Arrests)	1,488	0.915	-0.203	(0.193)	-0.3%

*** $p<0.001$, ** $p<0.01$, * $p<0.05$, + $p<0.1$
This table shows the coefficients from bivariate regressions of different arrest characteristics (measured at the officer level) on officer effects or arrest propensity, $\hat{\theta}_i$, with robust standard errors in parentheses. Officer effects are calculated using a training sample of data of 24 months covering 7/2014-6/2015 and 11/2017-10/2018, and is restricted to officers with over 500 call responses in this period. The officer level outcomes are calculated for the intervening data period from 7/2015-10/2017, depending on data coverage for the different characteristics (See Online Data Appendix A5 for more information on measure construction and data sources). Each outcome measure draws on information for an officer's total arrests and is not restricted to outcomes of dispatched 9-1-1 calls. The denominator of each measure is in parentheses of the outcome title, if relevant. The effect sizes of "Change 1 S.D. Officer Effect" refers to a percent change in the regression outcome due to a 1 standard deviation increase in the Officer Effect distribution, relative to the mean. Baseline characteristics correlate the Officer Effects measured in the training sample to relevant balance or check measures in the test sample. Column (1) measures the correlation to Officer Effects in the test sample (using a 500 call observation restriction in the test period). Columns (2), (3) and (4) measure the correlations between the training sample Officer Effect and unconditional arrest rates, total arrests and total call observations in the test sample. Columns (5) and (6) are measures of average monthly crime rates faced by each officer during the test sample; these are calculated as person-level weighted averages of division crime offense reports, based on the division in which an officer responds to calls during each month. Column (7) shows of the proportion of test sample arrests that result in a non-shooting use of force incident, where physical force was used or a civilian was injured, the civilian did not resist the officer and did not have a weapon and the officer was not injured. Columns (8) and (9) display the proportion of total arrests that are either felonies or misdemeanors, where these categories are not mutually exclusive and 46% of arrests are not coded as felonies or misdemeanors. Column (10) uses jail records to construct the proportion of test sample arrests that result in a jail booking, where the remainder of arrests are either Class C Misdemeanors that are not booked at all (citation issued), or arrests that involve a transfer to another type of facility (including medical), or another agency. Column (11) measures the proportion of each officer's jail admissions in the test sample that resulted in jail release because the booked individual was not able to be charged with a higher level offense than a Class C Misdemeanor; jail time for this class of low-level offenses is not permitted in Texas. Column (12) measures the proportion of test sample arrests that result in a dismissed/non-conviction charge and Column (13) measures the test sample conviction arrest rate. These rates are constructed by linking non-Class C Misdemeanor arrest records that include arrestee name and charge type ($\approx$20% of arrests) to court records using a fuzzy match by name and offense date (match rate of $\approx$ 89%). Dismissal and conviction rates are calculated among the linked arrest and court record sample.

Figure 4: Race Disparity in Arrests

Figure 4.A: Race Share of Arrests, Across Officer Arrest Propensity Deciles

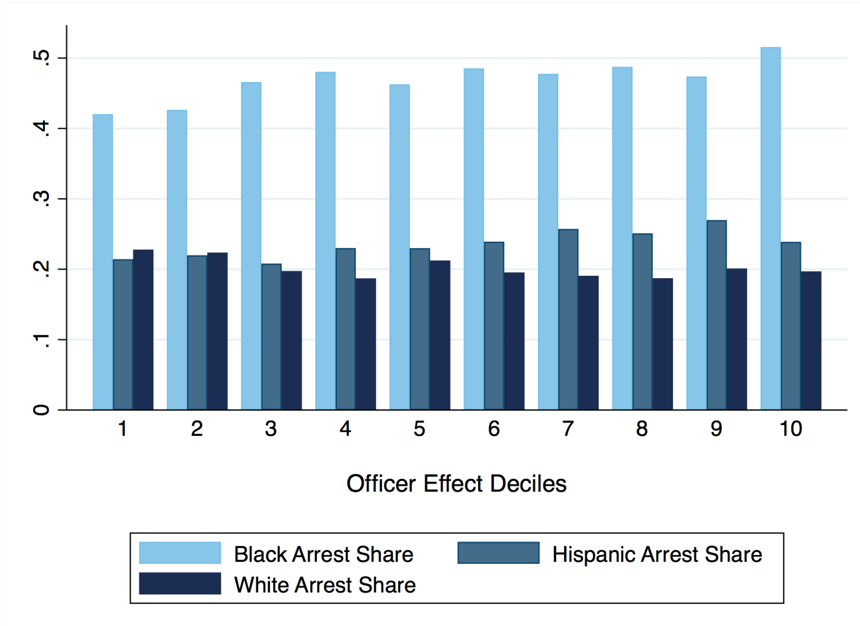
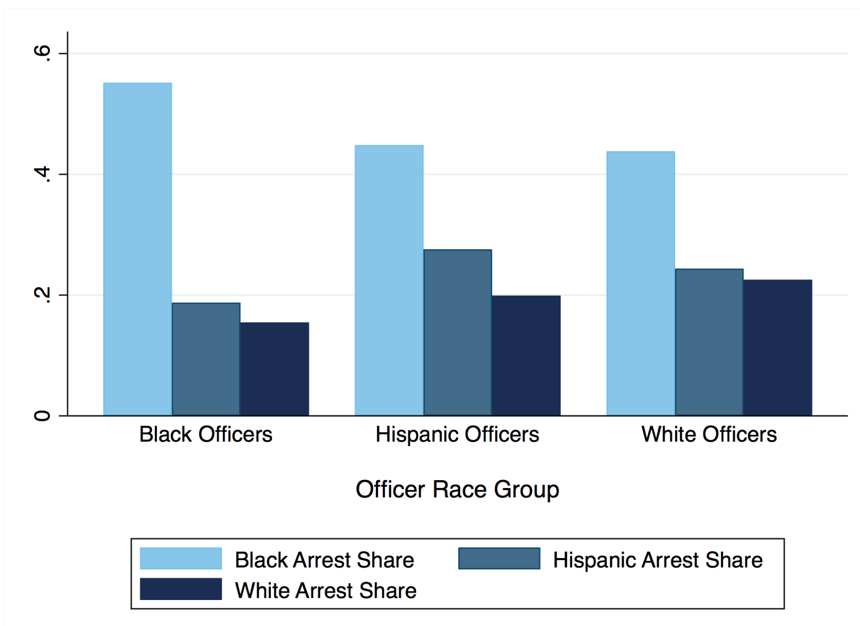


Figure 4.B: Race Share of Arrests, Across Officer Race



Each figure plots relative race shares of arrests (e.g. the proportion of Black arrests relative to total arrests) for different officer groups. These group rates are averages of race shares of arrests for each officer in the sample. The shares are descriptive means, they are unadjusted for covariates or characteristics of incidents or arrests. Race information is available for $\approx 90\%$ of arrests in the sample

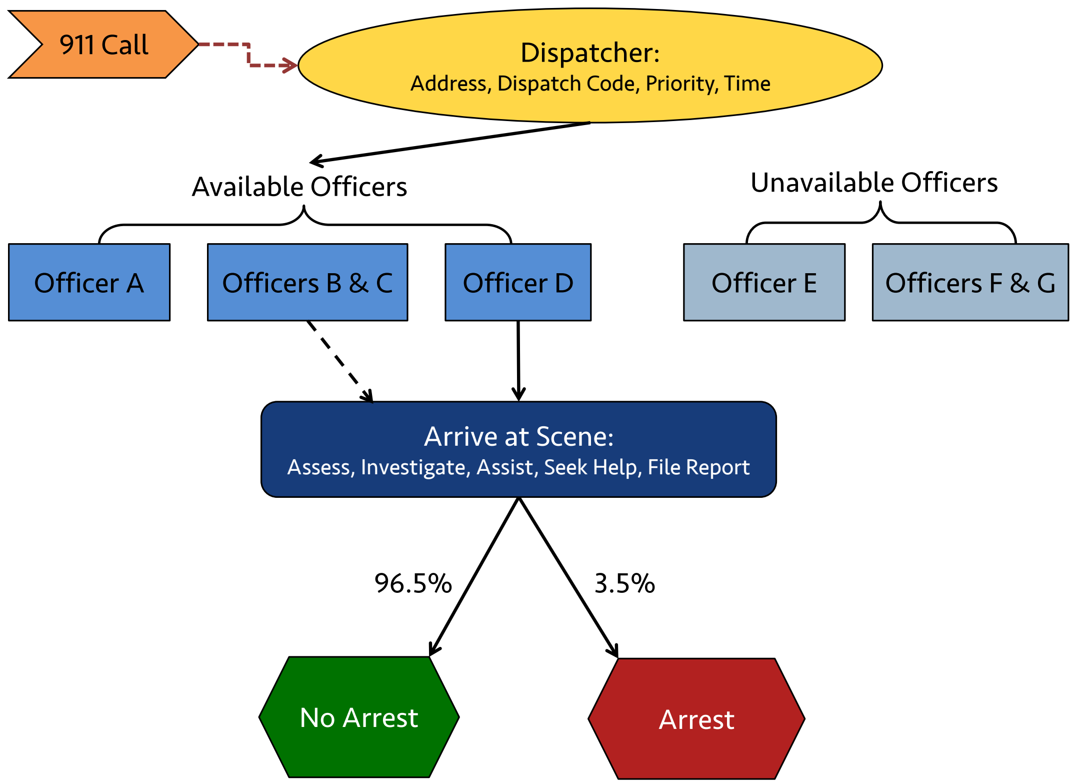
Table 5: Regression Adjusted Test of Taste-Based Racial Bias

Outcome: Arrestee Race Officer Effect	(1) Black Arrestee	(2) Hispanic Arrestee	(3) White Arrestee
A. Deciles of Total Arrest Propensity			
2nd Decile	0.003***	0.001***	0.001***
	(0.000)	(0.000)	(0.000)
3rd Decile	0.004***	0.002***	0.002***
	(0.000)	(0.000)	(0.000)
4th Decile	0.006***	0.003***	0.003***
	(0.000)	(0.000)	(0.000)
5th Decile	0.007***	0.003***	0.004***
	(0.000)	(0.000)	(0.000)
6th Decile	0.008***	0.004***	0.004***
	(0.000)	(0.000)	(0.000)
7th Decile	0.009***	0.005***	0.005***
	(0.000)	(0.000)	(0.000)
8th Decile	0.011***	0.006***	0.005***
	(0.000)	(0.000)	(0.000)
9th Decile	0.014***	0.006***	0.007***
	(0.000)	(0.000)	(0.000)
10th Decile	0.021***	0.008***	0.010***
	(0.001)	(0.000)	(0.001)
Observations	1,544	1,544	1,544
R-squared	0.627	0.469	0.458
Arrestee Race Mean	0.016	0.008	0.008
Change in Officer Dispersion (S.D.)			
from Fixing Total Arrest Propensity	-32.9%	-22.2%	-21.9%
Reject Null Hypothesis of no bias?	No	No	No
B. Officer Race			
Black	-0.002***	-0.000*	-0.001*
	(0.000)	(0.000)	(0.000)
Hispanic	-0.000	-0.000	0.000
	(0.001)	(0.000)	(0.000)
Observations	1,542	1,542	1,542
R-squared	0.010	0.004	0.008
Arrestee Race Mean	0.016	0.008	0.008
Change in Officer Dispersion (S.D.)			
from Fixing Race	-0.48%	-0.19%	-0.40%
Reject Null Hypothesis of no bias?	No	No	No

*** $p<0.001$, ** $p<0.01$, * $p<0.05$, + $p<0.1$
This table shows the results of the regression adjusted racial bias test, across officer total arrest propensity groups and officer race groups. The outcome in each regression is an officer effect, $\hat{\theta}_{i,r}$, obtained from a predictive model of unconditional arrestee race outcomes using the model framework described in Section 3. An unconditional arrestee race outcome is set equal to 1 if a call resulted in an arrest of an individual of a particular race and 0 otherwise. The racial bias test measures whether the relative rank order of officer group coefficients, either the decile of the officer arrest propensity distribution or officer race (displayed as regressors above), is the same for all arrestee races. Robust standard errors are in parentheses.

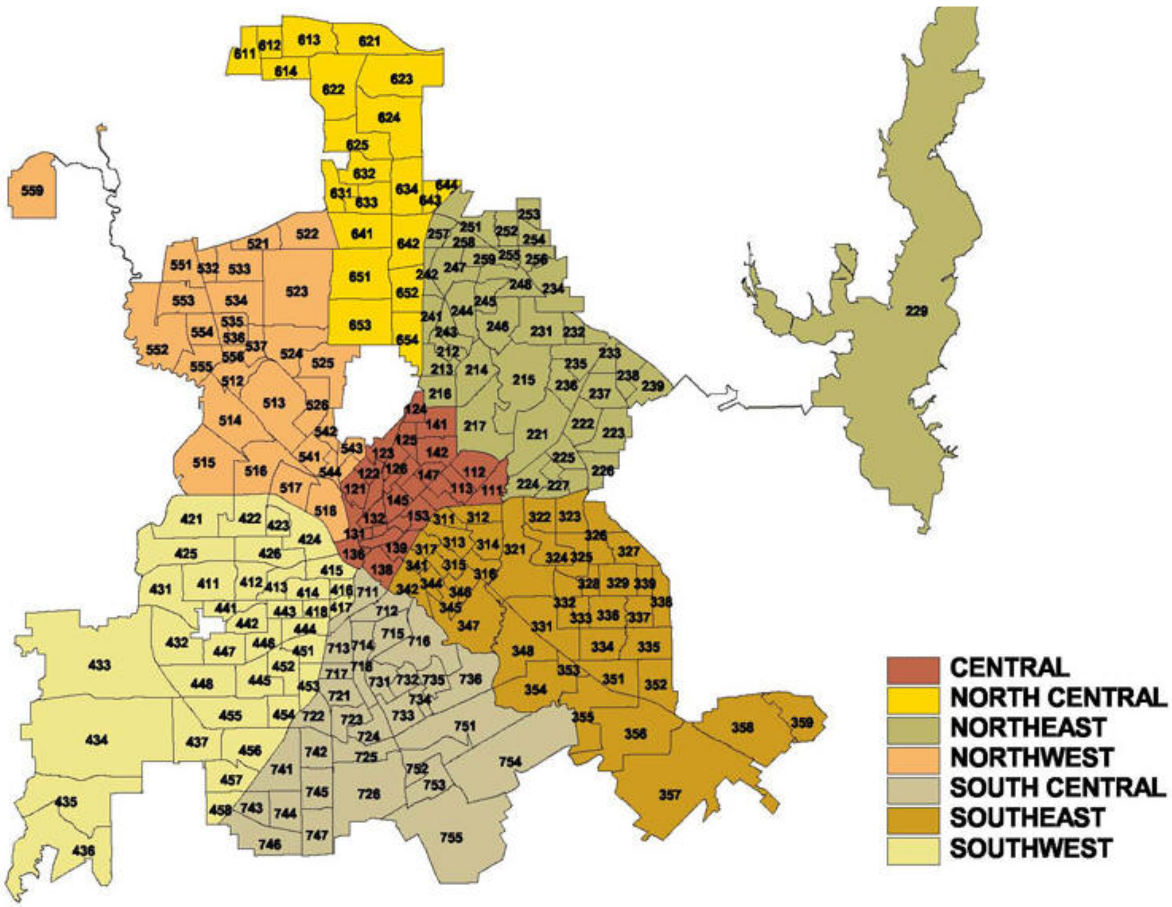
A1 Appendix Tables and Figures

Figure A1: Steps involved in a Response to an Offense Reported through a 9-1-1 Call



This figure displays an outline of a call incident response path at the Dallas Police Department. Information on call response protocols was obtained through conversations with officers and dispatchers at the department.

Figure A2: Police Beats and Police Divisions in Dallas, TX



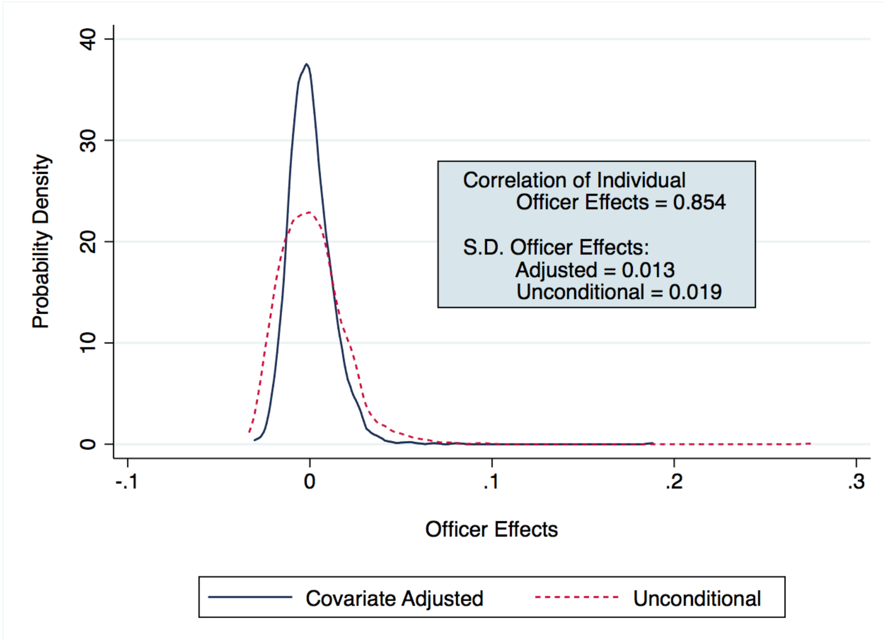
This figure shows a map of the 234 police beats contained in the 7 police divisions in Dallas. Police sectors are geographic units that are collections of beats within police divisions (35 total sectors). Map obtained from the North Dallas Neighborhood Alliance: `http://www.ndna-tx.org/crimeWatch/dallasPolice/DivMap.aspx`.

Table A1: Summary Statistics, Officer Sorting Robustness Samples

	Analysis Sample		*Low Availability*		*High Urgency*	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
Total Observations	3,425,027		1,687,415		1,992,561	
Total Call Responses	1,987,467		983,113		1,090,698	
Total Officers	1,544		1,509		1,542	
Primary Outcome: Arrest	0.036	(0.187)	0.036	(0.186)	0.036	(0.187)
Black Arrestee	0.016	(0.126)	0.016	(0.125)	0.016	(0.125)
Hispanic Arrestee	0.008	(0.086)	0.007	(0.084)	0.008	(0.087)
White Arrestee	0.008	(0.091)	0.009	(0.093)	0.009	(0.093)
Call Characteristics						
Time to Dispatch (Minutes)	14.56	(19.65)	14.29	(19.28)	3.20	(3.33)
Proportion of Officers Unavailable	0.280	(0.153)	0.400	(0.091)	0.282	(0.150)
Number of Responding Officers	1.97	(0.682)	1.96	(0.679)	2.07	(0.697)
Call Priority						
High Priority	0.647	(0.478)	0.642	(0.479)	0.739	(0.439)
Low Priority	0.353	(0.478)	0.358	(0.479)	0.261	(0.439)
Dispatch Code Type						
Criminal Assault	0.003	(0.059)	0.003	(0.058)	0.005	(0.068)
Armed Encounter/Active Shooter	0.013	(0.114)	0.014	(0.116)	0.020	(0.138)
Robbery	0.018	(0.133)	0.018	(0.132)	0.022	(0.146)
Burglary of Business	0.040	(0.196)	0.038	(0.190)	0.039	(0.194)
Burglary of Vehicle	0.020	(0.141)	0.018	(0.134)	0.015	(0.123)
Burglary of Residence	0.056	(0.230)	0.055	(0.229)	0.058	(0.234)
Unauthorized Use of Vehicle	0.010	(0.097)	0.009	(0.094)	0.005	(0.074)
Theft	0.012	(0.110)	0.012	(0.110)	0.008	(0.088)
Criminal Mischief	0.114	(0.318)	0.116	(0.321)	0.132	(0.338)
Major Disturbance	0.299	(0.458)	0.302	(0.459)	0.305	(0.460)
Injured Person	0.005	(0.074)	0.006	(0.075)	0.003	(0.058)
Accident	0.119	(0.323)	0.115	(0.319)	0.113	(0.316)
Other	0.290	(0.454)	0.294	(0.455)	0.275	(0.447)
Time of Day						
Day (8am-4pm)	0.335	(0.472)	0.338	(0.473)	0.321	(0.467)
Evening (4pm-12am)	0.423	(0.494)	0.429	(0.495)	0.418	(0.493)
Overnight (12am-8am)	0.242	(0.428)	0.233	(0.423)	0.261	(0.439)
Officer Characteristics						
Total Calls	2,548.7	(947.1)	1,271.0	(453.9)	1,488.7	(556.4)
Trainee	0.067	(0.250)	0.065	(0.247)	0.066	(0.249)
Sergeant	0.009	(0.095)	0.009	(0.095)	0.009	(0.096)
Salary ($10,000s)	57.30	(10.60)	57.06	(10.476)	57.14	(10.532)
Years of Experience	11.66	(8.48)	11.45	(8.350)	11.53	(8.422)
Age	37.23	(9.70)	37.04	(9.631)	37.08	(9.660)
Female	0.142	(0.349)	0.143	(0.350)	0.141	(0.348)
Black	0.260	(0.439)	0.262	(0.440)	0.257	(0.437)
Hispanic	0.213	(0.410)	0.215	(0.410)	0.214	(0.410)
White	0.484	(0.500)	0.480	(0.500)	0.486	(0.500)

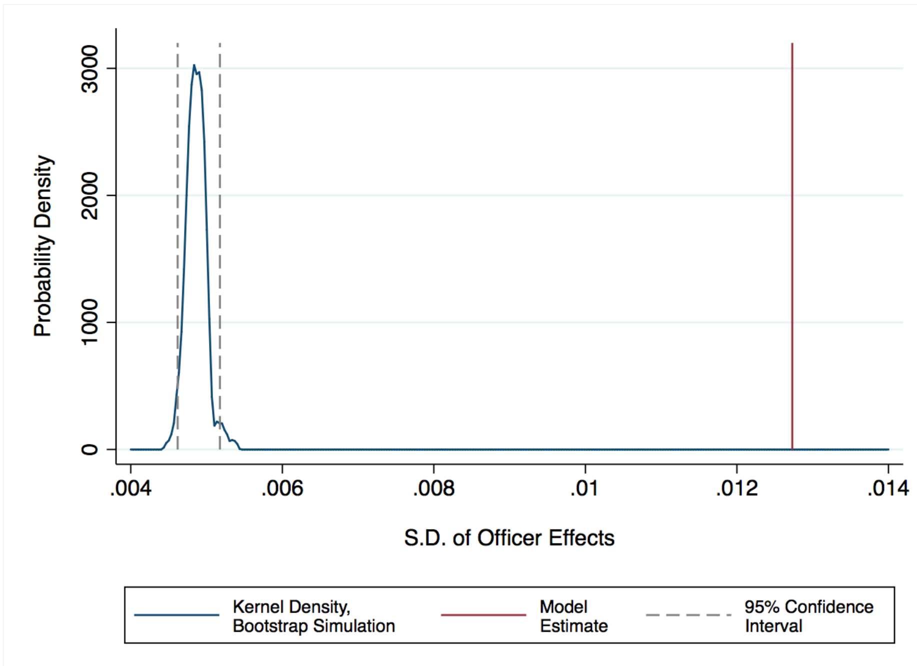
These tables display summary statistics for covariates used in analysis. The first column, "Analysis Sample", summarizes the primary analysis sample. The "Low Availability" sub-sample is determined by taking the set of observations where a greater proportion of officers are unavailable because they are responding to other offenses at the time an offense is dispatched, split at the median within broad patrol shift cells (8 Hour Shift*Day of the Week*Division Group). The "High Urgency" sub-sample consists of observations with a shorter time between when a call is received and when it is dispatched, split at the median within broad patrol shift cells (8 Hour Shift*Day of the Week*Division Group). The analysis is restricted to officers with at least 500 observations in the sub-samples.

Figure A3: Covariate-Adjusted vs. Unconditional Officer Effects



This figure plots the Officer Effects estimates relative to centered unconditional Officer Effects which are not adjusted for any of the covariates in the baseline model. This figure is similar to Figure 3.A which excludes "Officer Decision Controls", X_{kt} and ϕ_g and different car responding officers, or call characteristics that an officer may choose at the level of a call response. This figure additionally excludes shifts and same-car partners, both of which are determined prior to a call event.

Figure A4: Bootstrap Benchmark Test: Distribution of Results under Assumption of No "True" Officer Effects (or all Officer Effects are Jointly Zero)



This graph shows the residual bootstrap test distribution for the standard deviation of the officer effect distribution as well as the relative proportion of R^2 explained by officer fixed effects. Each bootstrap iteration is obtained as follows: (1) Residuals and predicted outcomes are obtained from a first stage model that does not include Officer FE (under the null hypothesis that these variables are jointly zero), (2) Estimated residuals are assigned a wild bootstrap weight of $w \in \{1, -1\}$ with equal probability that is constant within dispatched call response, and these residuals are added to the predicted outcomes from (1), (3) Using these simulated outcome variables, the full model, including Officer FE, is estimated to obtain each statistic of interest. Post-estimation Empirical Bayes' adjustments are made to the estimates after each iteration. Each test is based on 100 bootstrap replications.

I have also conducted this test imposing the restriction that $\tilde{Arrest}_b$ is binary in each iteration. To do this, I set the highest values of the outcome variable equal to one such that the mean of $\tilde{Arrest}_b$ equals the mean of $Arrest$ (approximately the top decile given an arrest mean of 10% in the sample. The results of this bootstrap test are similar and are available on request.

Table A2: Robustness Specification Tests

	(1) Base Model	(2) Add Reporting Area Fixed Effects	(3) Add Sector*Month Fixed Effects	(4) Add Individual Shift Effects	(5) Add Full Set of Dispatch Codes	(6) Crime Offense Report Sub-sample, N>100	(7) First Stage Fixed Effects: >2000 N	(8) First Stage Fixed Effects: Weighted by N	(9) First Stage Fixed Effects: Unadjusted
Total Officers	1,544	1,544	1,544	1,544	1,544	1,467	839	1,544	1,544
Total Observations	3,206,609	3,154,803	3,205,988	3,206,605	3,206,609	294,140	2,215,237	3,206,609	3,206,609
Mean of Outcome	0.034	0.034	0.034	0.034	0.034	0.088	0.032	0.034	0.034
S.D. of Outcome	0.182	0.182	0.182	0.182	0.182	0.283	0.177	0.182	0.182
Distribution of Officer Effects									
S.D. of Officer Effect	0.013	0.012	0.013	0.013	0.011	0.025	0.011	0.012	0.013
% Change: 1 S.D. Increase in Officer Effect	37.2%	36.0%	37.6%	38.2%	31.8%	28.9%	33.3%	35.8%	38.8%
Gap: 10th - 90th Percentile in Officer Effect	0.029	0.028	0.029	0.029	0.025	0.062	0.025	0.027	0.029
% Change: 10th - 90th Percentile in Officer Effect	83.9%	82.3%	83.4%	86.0%	71.9%	70.7%	79.1%	79.6%	85.1%
Contribution of Officer Effects									
Total R-2	0.074	0.084	0.073	0.089	0.129	0.232	0.073	0.074	0.074
Total Adjusted R-2	0.043	0.053	0.041	0.051	0.100	0.186	0.041	0.043	0.043
Additional R-2 from Officer Effects	0.051	0.052	0.051	0.050	0.051	0.056	0.050	0.051	0.051
Additional Adjusted R-2 from Officer Effects	0.023	0.024	0.023	0.022	0.024	0.030	0.021	0.023	0.023
% of R-2 from Officer Effects	69.4%	61.7%	70.9%	56.8%	39.2%	24.3%	68.5%	69.4%	69.4%
% of Adj. R-2 from Officer Effects	53.1%	44.4%	55.5%	43.2%	23.8%	16.4%	52.1%	53.1%	53.1%
Correlation to Arrest Officer Effect									
Full Samples		0.995	0.999	0.998	0.967	0.492	0.989	0.999	0.999

This table summarizes the analysis results across robustness specifications. Column (1) replicates the results from the primary specification. In column (2), police beat FE, ϕ_g, are replaced with reporting area FE, a finer geographic unit. In column (3), police beat FE, ϕ_g, are replaced with Sector*Month FE, which interact the 35 geographic police sectors with month indicators. Column (4) includes individual shift indicators or Date*Division*8 hour Shift fixed effects rather than shifts aggregated by month. Column (5) inserts the full set of 85 dispatch codes, rather than the broader 22 dispatch codes used in the preferred specification. Column (6) estimates Officer Effects from a sub-sample of calls that resulted in an incident offense report record, the sample that was the primary focus of analysis in earlier drafts of this project. This sample includes additional information on complainant demographics and location codes, but represents about a 10th of all dispatched calls (corresponding officer observation restriction is 100 observations in this sample). Column (7) reports the dispersion in unadjusted officer fixed effects from the first stage, using a sample restricted to officers with more than 2000 observations. Column (8) calculates the dispersion metric as a standard deviation of unadjusted officer fixed effects from the first stage that is weighted by the number of observations for each officer. Column (9) reports dispersion in the unadjusted officer fixed effect estimates. Correlations shown in the bottom panel display relationships between estimated Officer Effects from the different specifications.

Table A3: Officer Effects and Arrest Characteristics, including Officer Demographic Controls

Outcomes	Observations	Mean	Coefficient	Standard Error	*Change 1 S.D. Officer Effect*
A. Baseline Characteristics					
(1) Officer Effect, Test Sample	1,255	0.000	0.556***	(0.032)	19.4%
(2) Unconditional Arrest Rate	1,484	0.039	0.755***	(0.058)	24.4%
(3) Total Arrests	1,484	49.99	870.0***	(69.38)	24.9%
(4) Total Observations	1,484	1,354	-1,055	(1,328)	-2.4%
(5) Violent Crime (100,000 Residents)	1,484	81.82	31.72	(37.61)	0.5%
(6) Property Crime (100,000 Residents)	1,484	298.4	336.8*	(151.1)	1.4%
B. Arrest Characteristics					
(7) Use of Force Arrests (Total Arrests)	1,412	0.017	0.170*	(0.086)	12.6%
(8) Felony Arrests (Total Arrests)	1,502	0.149	-0.101	(0.213)	-0.9%
(9) Misdemeanor Arrests (Total Arrests)	1,502	0.406	1.313***	(0.365)	4.1%
(10) Jail Bookings (Total Arrests)	1,459	0.360	1.626***	(0.392)	5.7%
(11) Class C Misdemeanor Jail Releases (Jail Bookings)	1,460	0.151	0.660*	(0.262)	5.5%
(12) Dismissed/Non-conviction Arrests (Matched Arrests)	1,477	0.170	0.118	(0.278)	0.9%
(13) Conviction Arrests (Matched Arrests)	1,477	0.915	-0.296	(0.191)	-0.4%

*** $p<0.001$, ** $p<0.01$, * $p<0.05$, + $p<0.1$

This table replicates the analysis in Table 4, including additional officer level demographic covariates of race, gender, age, experience, experience squared, trainee status and sergeant status. The table shows the coefficients from bivariate regressions of different arrest characteristics (measured at the officer level) on officer effects or arrest propensity, $\hat{\theta}_i$, with robust standard errors in parentheses. Officer effects are calculated using a training sample of data of 24 months covering 7/2014-6/2015 and 11/2017-10/2018, and is restricted to officers with over 500 call responses in this period. The officer level outcomes are calculated for the intervening data period from 7/2015-10/2017, depending on data coverage for the different characteristics (See Online Data Appendix A5 for more information on measure construction and data sources). Each outcome measure draws on information for an officer's total arrests and is not restricted to outcomes of dispatched 9-1-1 calls. The denominator of each measure is in parentheses of the outcome title, if relevant. The effect sizes of "Change 1 S.D. Officer Effect" refers to a percent change in the regression outcome due to a 1 standard deviation increase in the Officer Effect distribution, relative to the mean. Baseline characteristics correlate the Officer Effects measured in the training sample to relevant balance or check measures in the test sample. Column (1) measures the correlation to Officer Effects in the test sample (using a 500 call observation restriction in the test period). Columns (2), (3) and (4) measure the correlations between the training sample Officer Effect and unconditional arrest rates, total arrests and total call observations in the test sample. Columns (5) and (6) are measures of average monthly crime rates faced by each officer during the test sample; these are calculated as person-level weighted averages of division crime offense reports, based on the division in which an officer responds to calls during each month. Column (7) shows of the proportion of test sample arrests that result in a non-shooting use of force incident, where physical force was used or a civilian was injured, the civilian did not resist the officer and did not have a weapon and the officer was not injured. Columns (8) and (9) display the proportion of total arrests that are either felonies or misdemeanors, where these categories are not mutually exclusive and 46% of arrests are not coded as felonies or misdemeanors. Column (10) uses jail records to construct the proportion of test sample arrests that result in a jail booking, where the remainder of arrests are either Class C Misdemeanors that are not booked at all (citation issued), or arrests that involve a transfer to another type of facility (including medical), or another agency. Column (11) measures the proportion of each officer's jail admissions in the test sample that resulted in jail release because the booked individual was not able to be charged with a higher level offense than a Class C Misdemeanor; jail time for this class of low-level offenses is not permitted in Texas. Column (12) measures the proportion of test sample arrests that result in a dismissed/non-conviction charge and Column (13) measures the test sample conviction arrest rate. These rates are constructed by linking non-Class C Misdemeanor arrest records that include arrestee name and charge type (≈20% of arrests) to court records using a fuzzy match by name and offense date (match rate of ≈ 90%). Dismissal and conviction rates are calculated among the linked arrest and court record sample.

Table A4: Regression Adjusted Test of Taste-Based Racial Bias, with Officer Demographic Controls

	(4)	(5)	(6)
Outcome: Arrestee Race Officer Effect	Black Arrestee	Hispanic Arrestee	White Arrestee
A. Deciles of Total Arrest Propensity			
2nd Decile	0.003***	0.001***	0.001***
	(0.000)	(0.000)	(0.000)
3rd Decile	0.004***	0.002***	0.002***
	(0.000)	(0.000)	(0.000)
4th Decile	0.005***	0.003***	0.003***
	(0.000)	(0.000)	(0.000)
5th Decile	0.006***	0.003***	0.004***
	(0.000)	(0.000)	(0.000)
6th Decile	0.008***	0.004***	0.004***
	(0.000)	(0.000)	(0.000)
7th Decile	0.009***	0.005***	0.005***
	(0.000)	(0.000)	(0.000)
8th Decile	0.011***	0.006***	0.005***
	(0.000)	(0.000)	(0.000)
9th Decile	0.014***	0.006***	0.007***
	(0.000)	(0.000)	(0.000)
10th Decile	0.021***	0.009***	0.011***
	(0.001)	(0.000)	(0.001)
Observations	1,542	1,542	1,542
R-squared	0.628	0.471	0.465
Arrestee Race Mean	0.016	0.008	0.008
Change in Officer Dispersion (S.D.)			
from Fixing Total Arrest Propensity	-33.1%	-22.1%	-21.6%
Reject Null Hypothesis of no bias?	No	No	No
B. Officer Race			
Black	-0.002***	-0.000*	-0.001*
	(0.000)	(0.000)	(0.000)
Hispanic	-0.000	-0.000	0.000
	(0.001)	(0.000)	(0.000)
Observations	1,542	1,542	1,542
R-squared	0.010	0.004	0.008
Arrestee Race Mean	0.016	0.008	0.008
Change in Officer Dispersion (S.D.)			
from Fixing Race	-0.4%	-0.1%	-0.3%
Reject Null Hypothesis of no bias?	No	No	No

*** $p<0.001$, ** $p<0.01$, * $p<0.05$, + $p<0.1$
This table shows the results of the regression adjusted racial bias test, across officer total arrest propensity groups and officer race groups. The outcome in each regression is an officer effect, $\hat{\theta}_{i,r}$, obtained from a predictive model of unconditional arrestee race outcomes using the model framework described in Section 3. An unconditional arrestee race outcome is set equal to 1 if a call resulted in an arrest of an individual of a particular race and 0 otherwise. The racial bias test measures whether the relative rank order of officer group coefficients, either the decile of the officer arrest propensity distribution or officer race (displayed as regressors above), is the same for all arrestee races. Robust standard errors are in parentheses. This output also includes controls for officer race, gender, age, experience, experience squared, trainee and sergeant status.

Figure A5: Race-Specific Officer Effects Distribution by Officer Arrest Propensity

Figure A5.A: Black Arrest Outcome

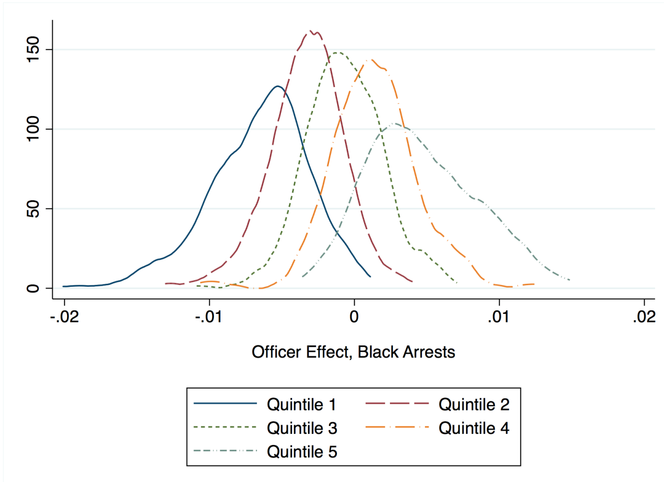


Figure A5.B: Hispanic Arrest Outcome

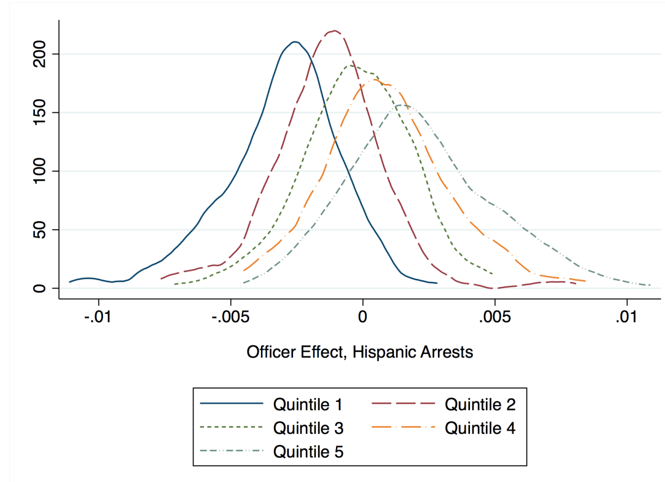
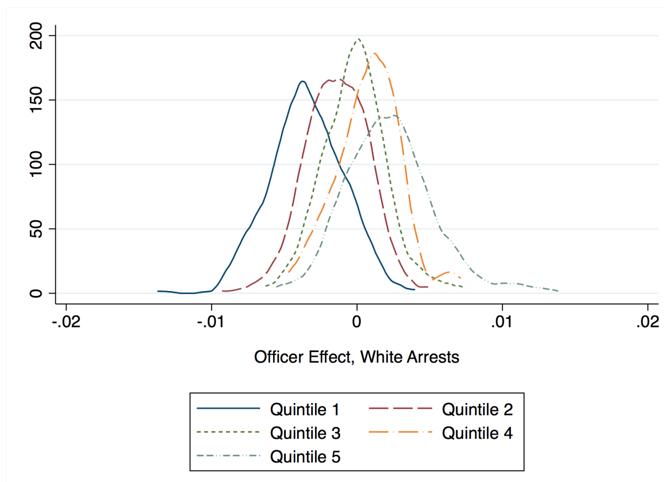


Figure A5.C: White Arrest Outcome



The graphs above plot the distribution of Officer Effects for different unconditional race-specific arrest outcomes of calls for service. For example, the "Black Arrest Outcome" is equal to 1 if a call for service results in an arrest of a Black civilian and 0 otherwise. Officer Effects for each outcome are estimated using the fully specified model covariates described in Section 3. These Officer Effects for each race-specific outcome are plotted separately for quintiles of the overall Officer Effect distribution measured for the total arrest outcome. The racial bias test is not rejected when the order of these distributions across quintile groups significantly differs for different race-specific arrest outcomes.

Figure A6: Race-Specific Officer Effects Distribution by Officer Race

Figure A6.A: Black Arrest Outcome

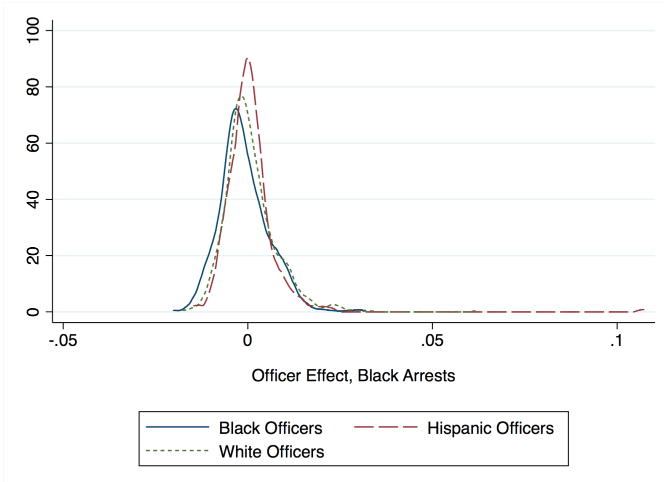


Figure A6.B: Hispanic Arrest Outcome

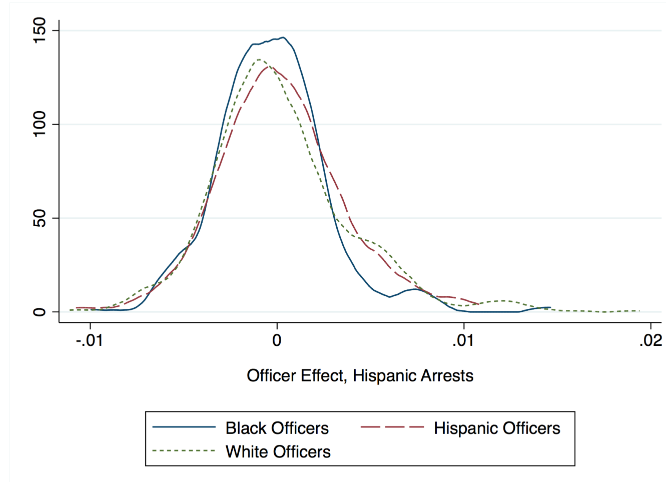
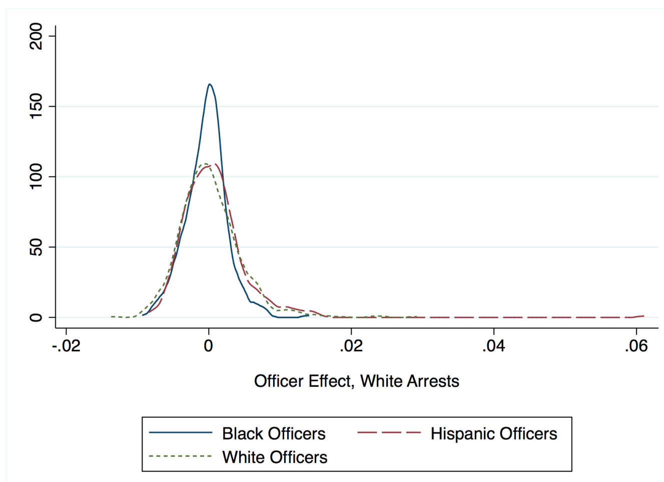


Figure A6.C: White Arrest Outcome



The graphs above plot the distribution of Officer Effects for different unconditional race-specific arrest outcomes of calls for service. For example, the "Black Arrest Outcome" is equal to 1 if a call for service results in an arrest of a Black civilian and 0 otherwise. Officer Effects for each outcome are estimated using the fully specified model covariates described in Section 3. These Officer Effects for each race-specific outcome are plotted separately for different officer race groups. The racial bias test is not rejected when the order of these distributions across officer race groups significantly differs for different race-specific arrest outcomes.

A2 Coefficients in the First Stage of Model (Online Appendix)

In the body of the paper, I restrict attention to aspects of officer effects because this paper focuses on estimating differences in officer arrest behavior, $\hat{\theta}_i$, and the importance of officers in predicting arrests. This appendix discusses other components of the arrest prediction model.

Table A5: Covariate Coefficients in First Stage of Arrest Model

Dispatch Codes			
Criminal Assault, High Priority	0.119***	Theft, High Priority	0.024***
	(0.006)		(0.002)
Armed Encounter/Active Shooter, High Priority	0.035***	Theft, Low Priority	-0.012***
	(0.002)		(0.001)
Robbery, High Priority	0.007***	Criminal Mischief, High Priority	-0.019***
	(0.001)		(0.001)
Burglary of Business, High Priority	0.023***	Criminal Mischief, Low Priority	0.003*
	(0.004)		(0.001)
Burglary of Business, Low Priority	-0.034***	Major Disturbance, High Priority	0.002**
	(0.001)		(0.001)
Burglary of Residence, High Priority	-0.007***	Accident, High Priority	-0.007***
	(0.001)		(0.001)
Burglary of Residence, Low Priority	-0.026***	Accident, Low Priority	-0.020***
	(0.001)		(0.001)
Burglary of Vehicle, High Priority	0.034***	Injured Person, High Priority	0.004
	(0.003)		(0.004)
Burglary of Vehicle, Low Priority	-0.022***	Injured Person, Low Priority	0.004+
	(0.001)		(0.002)
Unauthorized Use of Vehicle, High Priority	0.058***	Other, High Priority	0.001
	(0.005)		(0.001)
Unauthorized Use of Vehicle, Low Priority	0.014***	*Omitted Category: Other, Low Priority*	
	(0.002)		
Call Urgency			
Time to Dispatch (Minutes)	-0.0006***		
	(0.00003)	Observations	3,206,609
Time to Dispatch (Minutes), Squared	0.00001***	R-Squared	0.074
	(0.0000)	Adjusted R-Squared	0.043
Proportion of Officers Available	0.011***	Arrest Outcome Mean	0.034
	(0.001)	Arrest Outcome Standard Deviation	0.182
Other Controls in Model:	First Officer and Group Co-responder fixed effects, Shift (Day-of-the-week*8-Hour Shift*Month*Year) fixed effects, Police Beat fixed effects, Hour fixed Effects		

*** $p<0.001$, ** $p<0.01$, * $p<0.05$, + $p<0.1$
This table shows the results of the first stage arrest model. Coefficients are shown for each covariate in the vector X_{kt}. Robust standard errors are clustered at the level of the focal officer, i, and the shift category, δ_{gt}.

Table A5 shows the first stage coefficients for incident characteristics, X_{kt}. First, the probability of an arrest is increasing in call severity or urgency, at a decreasing rate. This is shown by the "Time to Dispatch" variables that measure the number of minutes that lapse between when a call is made by the complainant and when an officer is dispatched to the scene. An increase of 10 minutes in this time gap decreases the likelihood that an arrest is made by 9% (relative to the average time gap). The average call in the data has a time difference of 14.5 minutes between the call and dispatch time, which corresponds to a decrease the likelihood of arrest by 14% relative to an instantaneously dispatched call.

The second set of incident controls in the model are dispatch codes. These variables are generally more positive for crimes that are more serious or have a higher priority level. The omitted category of minor incidents (other minor). A 1 standard deviation in the officer effect distribution is comparable to the difference in arrest likelihood moving from a "High Priority Theft" incident to a "Armed Encounter/Active Shooter." While hour of day fixed effects are not shown on this table, the coefficients show that arrests are more likely during the early hours of the overnight shift (12am-4am) and least likely in the pre-dawn hours of morning (5am-8am).

Overall, the incident context controls are important predictors of whether an incident results in an arrest and account for $\approx 15-25\%$ of the explainable variation in arrest outcomes.

A3 Empirical Bayes Shrinkage Estimates (Online Appendix)

As outlined in the text, the estimates of permanent officer arrest propensity are adjusted using Empirical Bayes techniques. These techniques are detailed in work by Morris (1983) and are commonly employed in the economics of education literature on teacher value added (e.g. Guarino et al., 2015; Koedel et al., 2015; Chetty et al., 2014; Kane and Staiger, 2008; Aaronson et al., 2007). In robustness checks in the paper, I show that the results do not substantively change when a number of alternate precision adjustments are used.

I observe sample estimates of officer arrest propensity, $\bar{r}_i$, which are derived from a first stage regression model. Each of these estimates is an approximation of a "true" officer arrest propensity, θ_i, though some officer estimates are derived from more observations and are thus more precise than others. Empirical Bayes techniques develop a "prior" distribution for the underlying distribution of θ_i that is estimated empirically from the data on all officers. The estimation constructs a weighted mean of the observational estimate and the "prior."

Each θ_i is assumed to be independent and identically distributed across G total officers. The underlying distribution of each $\bar{r}_i$ and the total distribution of θ_i across i are given by:

$$\begin{aligned} \bar{r}_i|\theta_i &\sim N(\theta_i, \frac{\sigma^2_{\varepsilon,i}}{N_i}) \\ \theta_i|\mu, \sigma^2_A &\sim N(0, \sigma^2_A) \end{aligned}$$

The mean of the distribution of θ_i is known to be 0 in this setting, given the normalization of the fixed effects in the model. Given a "prior" for the distribution of θ_i, the posterior distribution of $\theta_i|\bar{r}_i$ give the adjusted estimates of $\hat{\theta}^{EB}_i$ used in this paper:

$$\begin{aligned} \theta^{EB}_i|\bar{r}_i, \sigma^2_{\varepsilon,i}, \sigma^2_A &\sim N(B\bar{r}_i, B\frac{\sigma^2_{\varepsilon,i}}{N_i}) \\ where\ B &= \frac{\sigma^2_A}{\sigma^2_A + \frac{\sigma^2_{\varepsilon,i}}{N_i}} \end{aligned}$$

I derive estimates of officer arrest propensity, $\hat{\theta}^{EB}_i$, using the following steps:

1. Estimate the first stage of the model and calculate residuals, $\hat{r}_{ikgt}$, and their officer-level average, $\bar{r}_i$. I include all officer fixed effects in the first stage regression to allow for arbitrary correlations between responding officers and the other covariates in the

model to improve the estimation of the residuals, following Chetty et al. (2014).

$$\begin{aligned} Arrest_{ikgt} &= \theta_i + \theta_j + \pi X_{kt} + \delta_{gt} + \phi_g + \varepsilon_{ikgt} \\ \hat{r}_{ikgt} &= \hat{\theta}_i + \hat{\varepsilon}_{ikgt} \\ \bar{r}_i &= \frac{1}{N_i} \sum_{N_i} \hat{r}_{ikgt} \end{aligned}$$

2. Calculate individual variance estimates, $\hat{\sigma}^2_{\varepsilon,i}$ and solve for a sample analog of the prior variance of θ_i, $\hat{\sigma}^2_A$.

$$\begin{aligned} \hat{\sigma}^2_{\varepsilon,i} &= \frac{1}{N_i - 1} \sum_{N_i} (\hat{r}_{ikgt} - \bar{r}_i)^2 \\ \sigma^2_A &= E[r^2_{ikgt}] - E[\varepsilon^2_{ikgt}] \\ \hat{\sigma}^2_A &= \frac{1}{N - G - K} \sum_G \sum_{N_i} \hat{r}^2_{ikgt} - \frac{1}{N - G} \sum_G N_i \hat{\sigma}^2_{\varepsilon,i} \end{aligned}$$

with $N - G - K$ are the degrees of freedom in the first stage regression, given G officers and K regressors in the first stage model.[27]

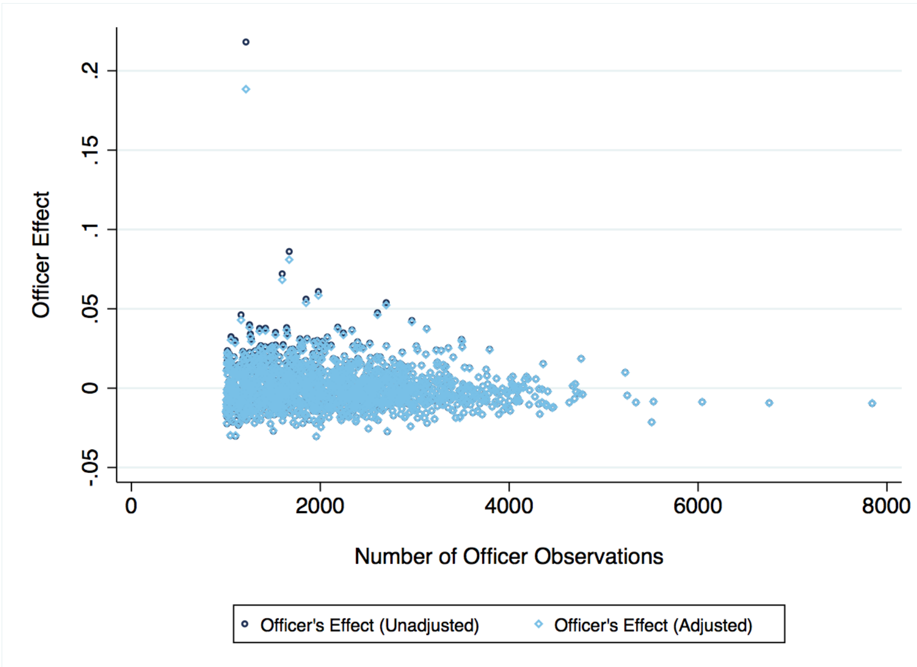
3. Calculate the posterior estimates $\hat{\theta}^{EB}_i$ by applying the shrinkage factor $\hat{B}$. The shrinkage factor is always less than 1 and is increasing in N_i and decreasing in $\hat{\sigma}^2_{\varepsilon,i}$. This factor gives higher weight to police officer arrest propensity estimates that are more precisely measured and shrinks less precise estimates toward 0, the center of the distribution.

$$\hat{\theta}^{EB}_i = \frac{\hat{\sigma}^2_A}{\hat{\sigma}^2_A + \frac{\hat{\sigma}^2_{\varepsilon,i}}{N_i}} \cdot \bar{r}_i$$

[27]Given that I absorb four sets of fixed effects in the model, θ_i, θ_{-i}, δ_{dt} and ϕ_g, K contains the number of group categories in the non-focal fixed effects. In practice, the degrees of freedom must also be adjusted for the number of omitted reference categories in the model, or the number of "mobility groups", M. The actual degrees of freedom used is $N - G - K + M$.

The following figure displays the relationship between the unadjusted and adjusted estimates of officer effects and the number of observations per officer:

Figure A7: Adjusted and Unadjusted Officer Effects



The above figure shows unadjusted officer fixed effects overlaid with the adjusted officer effects estimates used in this paper. The correlation between the adjusted and unadjusted estimates is 0.987.

A4 Economic Model for Racial Bias Test (Online Appendix)

In this section, I outline the model used to test for the presence of racial bias among officers. The model is adapted from the test of racial bias in Anwar and Fang (2006) to the setting of police calls for service.

There are two races of suspects in the illustrative model, $r \in \{M, W\}$. Officers are categorized into types $k \in K$, which in most specifications refer to an arrest propensity type or an officer race group. The model examines how race-specific policing outcomes vary across these types of officers.

When officers arrive to respond to an incident they observe whether a suspect is present and information about the suspect's race. The likelihood that a call response has a suspect of race r_s is given by ψ^{r_s}. Note that $\psi^M + \psi^W \leq 1$ because some call responses may not be associated with a crime that has a party that is at fault, such as an accidental injury. For each suspect race, r_s, π^{r_s} is the likelihood that a potential suspect is *guilty*. This probability will differ by suspect race if observable or unobservable characteristics of suspect criminality are correlated with race.

Next, the officer determines if there is sufficient evidence or basis for an arrest and decides whether to arrest the suspect. Officers make the decision to arrest a suspect based on observing a summarized signal of guilt, $s \in [0, 1]$, for potential suspects identified at the scene of a call response. This information may include evidence immediately available at the scene, cooperation of the victim, location of the incident, etc.

If a potential suspect is guilty, s is randomly drawn from the distribution $F_g^{r_s}(s)$, while if the individual is not guilty, s is randomly drawn from the distribution $F_n^{r_s}(s)$. These distributions are allowed to differ across suspect race, r_s, reflecting the fact that total information content in responses may differ for different suspect races.

Officers make a discrete choice in the model, whether or not to arrest a potential suspect. An officer will make an arrest if he determines that there is a minimum basis of guilt for an arrest.

Officers receive a benefit normalized to 1 if they arrest a guilty suspect. Officers face a cost of arrest of $t(r_s, k)$ that varies across officer group, k.

Officers choose whether or not to arrest a potential suspect to maximize

$$max\{P(G|s, r_s) - t(r_s, k), 0\}$$

The marginal probability of suspect guilt is given by Bayes' Rule and is strictly increasing in s:

$$P(G|r_s, s) = \frac{\pi^{r_s} f_g^{r_s}(s)}{\pi^{r_s} f_g^{r_s}(s) + (1 - \pi^{r_s}) f_n^{r_s}(s)}$$

Officers will make an arrest when $P(G|r_s, s) \geq t(r_s, k)$. As a result, it can be shown that officers will exert effort on arrests if the value of $s \geq s^*(r_s, k)$, where the threshold $s^*(r_s, k)$ satisfies $P(G|r_s, s^*(r_s, k)) = t(r_s, k)$. The effort signal threshold is monotonically increasing in officer costs.

The distributions $f_g^{r_s}(s)$ and $f_n^{r_s}(s)$ have the following properties:

- Both are defined over the full support of $s \in [0,1]$
- Monotone Likelihood Ratio Property: $\frac{f_g^{r_s}(s)}{f_n^{r_s}(s)}$ is strictly increasing in s. This implies that a higher s means an arrest is more likely to be feasible.
- Unbounded Likelihood Ratio: $\frac{f_g^{r_s}(s)}{f_n^{r_s}(s)} \to \infty$ as $s \to 1$. This implies that very high signals θ provide nearly certain information that an arrest is feasible.

In the data, we observe unconditional arrestee race outcomes, or the proportion of arrests of a particular race group relative to total calls for each officer group, k.

The race-specific arrest given call response rate is defined as:

$$A(r_s,k) = \psi^{r_s}[\pi^{r_s}(1 - F_g^{r_s}(s^*(r_s,k)) + (1-\pi^{r_s})(1 - F_n^{r_s}(s^*(r_s,k))]$$

where $A(r_s,k)$ is decreasing in $s^*(r_s,k)$ and $t(r_s,k)$.

The following definitions characterize officer race-specific costs:

1. Racial Bias: Officers are racially biased with respect to suspects if for some officer, k, $t(M,k) \neq t(W,k)$ or $c(M,k) \neq c(W,k)$.
2. Monolithic Behavior: Officers are not monolithic in their behavior if officer costs differ across officer type for a given suspect race. This will occur when $t(r_s,k) \neq t(r_s,k')$ or $c(r_s,k) \neq c(r_s,k')$.
3. Statistical Discrimination: Assume $t(M,k) = t(W,k)$ and $c(M,k) = c(W,k)$, or officers are not racially biased. Then type k officers will exhibit statistical discrimination if $s^*(M,k) \neq s^*(W,k)$.

If officers are not racially biased and exhibit monolithic behavior across officer race, then $t(M,k) = t(M,k') = t(W,k) = t(W,k')$, $\forall k, k' \in K$). It follows that each observed race-specific arrest rate will be constant across officer groups, but that the arrest rates for different suspect races may differ if $s^*(M,k) \neq s^*(W,k)$, or there is statistical discrimination.

The test also allows officers to have differing total costs of effort that vary by officer group, or behave in a manner that is *non-monolithic*. The first half of this paper assesses whether individual officers differ in their arrest behavior when responding to similar incidents, which can be interpreted as evidence that individual officers are not monolithic in their behavior. If officers do not exhibit monolithic behavior but are also not prejudiced, then the ranking of race-specific arrest rates, across officer groups within suspect race will be independent of suspect race, or the same for each suspect race.

For example, allow k officers to have a higher cost of arrest effort than k' officers for any race of civilian, but allow both officer groups to be unbiased. Then:

$$
\begin{aligned}
t(M,k) > t(M,k') \quad &\& \quad t(W,k) > t(W,k') \\
t(M,k) = t(W,k) \quad &\& \quad t(M,k') = t(W,k') \\
s^*(M,k) > s^*(M,k') \quad &\& \quad s^*(W,k) > s^*(W,k') \\
A(M,k) < A(M,k') \quad &\& \quad A(W,k) < A(W,k')
\end{aligned}
$$

Or in this case, type k officers will be less likely than type k' officers to make arrests of any race group. In other words, the relative ranking of k and k' officers is the same for both suspect race groups.

Generally, the test proposed in this paper allows total arrest rates to differ across arrestee race by focusing attention on relative rankings of officer rates across type rather than total levels of officer rates. This feature allows officers to behave in a manner that is consistent with statistical discrimination and isolates officer behavioral patterns associated with taste-based racial bias. Statistical discrimination will occur in this model if total arrest rates for one suspect group are always higher than arrest rates for the other suspect group but the relative ranking of officer arrest rates is the same for both suspect groups. For example, it may be the case that Black and Hispanic suspects are more likely to have a criminal history and this causes the total signal threshold to be lower for incidents with minority suspects, $s^*(M,k) < s^*(W,k)\ \forall k \in K$.

Conversely, a reversal in the rank order of arrest rates across officer type for different suspect race groups will provide evidence of taste-based racial bias. This opposing rank order violates the null hypothesis of no racial bias among officers.

This is illustrated by the following stylized example:

$$
\begin{aligned}
t(M,k) > t(W,k) \quad &\& \quad t(W,k') > t(M,k') \\
\&\ t(W,k') = t(M,k) \quad &\& \quad t(W,k) = t(M,k') \\
s^*(M,k) > s^*(M,k') \quad &\& \quad s^*(W,k) < s^*(W,k') \\
A(M,k) < A(M,k') \quad &\& \quad A(W,k) > A(W,k')
\end{aligned}
$$

In this paper, I test for racial bias using two sets of officer group categories, officer arrest propensity groups, measured as deciles in the total arrest officer effect, $\hat{\theta}_i$ distribution, and officer race groups. I define arrestee race outcomes as equal to 1 if a call for service results in an arrest of an individual with a particular race and 0 otherwise. These outcomes are regressed on officer fixed effects using the full sample of dispatched calls and the fully specified model described in Section 3 in order to recover race-specific arrest officer effects, $\hat{\theta}_{i,r}$. I then conduct the rank order racial bias test by regressing $\hat{\theta}_{i,r}$ estimates on the officer group categories and checking whether the order of officer groups is the same for all arrestee race outcomes. The null hypothesis of no taste-based racial bias is rejected when one group, k, has a statistically significant higher arrest rate outcome than another group, k', for a given arrestee race, and this relationship is reversed and significant for a different arrestee race.

A5 Data Appendix (Online Appendix)

Several different data files were used for this project. This Appendix summarizes the decisions made in cleaning and constructing the data set used for this project.

Estimation of Officer Effects, θ_i

Dispatch Data The base analysis file consists of records of all events dispatched by DPD from 6/2014 to 10/2018, obtained through an Open Records Request to the city of Dallas. This data includes address, police beat, sector and division, dispatch code and time stamps for when the call occurred, was dispatched, officers arrived and the conclusion of the response. The data also includes records for each officer that responds to a dispatched call.

While the data is pulled from the call dispatch system, the data includes some call events that are officer-initiated rather than complainant-initiated. I clean the ≈ 120 dispatch codes into 22 groupings, to increase power and remove very small categories (the full set is included in a robustness specification). I clean the data to exclude calls listed as officer "Mark-Outs", or records where an officer initiate an investigation and then convey their location to the dispatcher, traffic stops, calls where officers respond to assist other officers in the field, fire related calls and other officer-initiated investigations. I also exclude calls in which the time between when the call was made and when the call was dispatched exceeds 1.5 hours.

I trim the sample to include only the set of officers that are first dispatched to a call (groups are often dispatched at the same time), but allow officers to arrive at the calls at different times. For computational purposes, I limit the sample to calls with 4 or fewer responders, a restriction that excludes approximately 1% of the data. I use this sample to calculate peer group identifiers for each officer on each call. The final restriction for the analysis data set limits the sample to call responses for officers with more than 1,000 observations.

I use the untrimmed data, including all officer-initiated responses and dispatched officers, to calculate the proportion of officers that are available relative to those observed working a particular 8 hour shift at the time of each call observation. An officer is designated as available if he is not observed responding to another call at the time of each observation. This "availability rate" measure is used both as a control in the model and to determine the "Low Availability" sample, or below median availability rate observations within a day of the week by division by 8 hour shift time slot.

Arrest Outcome Data I use four DPD data files to compile information on arrests: records of "Police Reported Incidents", "Police Arrests" and "Police Arrest Charges" accessed through the Dallas Open Data Portal, as well as an open records request for DPD arrest records over the sample time period. I use these files to create a liberal and comprehensive measure of whether any arrest occurred in association with a dispatched call. I use these files to code information on the arrest charge type, felony vs. misdemeanor, arrest date, arrestee names and arrestee race and gender demographics.

Officer Demographic Data Officer demographic information obtained through an open records request to the city of Dallas. Through this request, I acquired records of all police department employees from 2014 to the present that include officer names, badge number, job title, hire date, leave date (if applicable), ethnicity or race, gender, age and salary. I match the request records to the incident file using officer badge numbers and match officers by name if badge numbers are not available.

Because the officer request file includes employee title and badge number, I also use this information to exclude dispatch call observations with responding officers that are civilian police employees, as these incidents likely involved only a phone response and did not entail a physical patrol officer response to the scene.

Arrest Characteristics Measures

Measures of arrest characteristics are used in the second half of the paper to relate officer arrest propensity to other measures of police activity. A number of additional data sources are used to construct the different measures. All measures are calculated at the incident level in the sense that an arrest in the numerator or denominator corresponds to "any arrest for a particular incident" rather than at the person-arrest level which could include multiple arrests per incident. In this analysis, I relate officer arrest propensity estimates, $\hat{\theta}_i$, that are derived from a training data sample to arrest characteristics in a test sample. The training data set is always the first and last year of the data set or 7/2014-6/2015 and 11/2017-10/2018, while the test data set slightly differs in coverage for the different arrest characteristic data sets. I chose this bookended training data set to maximize the coverage of the data sets available for arrest characteristics, most of which end in the beginning of 2018.

Reported Crime Incident Data I use data on crimes reported to police or the "Police Incidents" file published through the Dallas Open Data Portal to estimate officer level crime rates. I first calculate counts of crimes at the police division by month level (per capita). Each incident includes UCR offense codes that are used to categorize offenses into property crimes, violent crimes, or other crimes. Violent crimes include aggravated assault, other assault, robbery, murder/manslaughter and weapons charges. Property crimes include burglary, theft, vehicle theft, embezzlement, fraud, forgery and counterfeiting. Population for each police division is calculated by assigning census tracts to a primary division and merging census tract level population data from 2012-2016 American Community Survey. I calculate officer level crime rates by averaging division month level crime rates across each officer's observed divisions and time periods in the data.

In earlier versions of this project, the analysis was focused on a base sample of "Police Incidents" or crime incident reports resulting from 9-1-1 calls. The current version of the project focuses on all dispatch calls, as officers may exercise some discretion about whether a report is written about an incident. Dispatch data also has the advantage of having a higher frequency and can include hundreds to thousands of observation per officer. Alternatively, an advantage of the crime incident data is that this data includes rich information on complainant demographics and incident location type. For this reason, I include a model estimated using the incident data as a robustness check.

Arrest Data: Charge Level and Demographics Misdemeanor and felony arrest information are taken from the combined arrest data sources described above. These are used to calculate officer-level misdemeanor or felony arrest shares (relative to total arrests). The denominator of total arrests includes arrest incidents where the charge type was not recorded, approximately 46% of arrests. Likewise, arrest shares by race group are also calculated using this data. The denominator for race arrest shares includes arrests with no recorded arrestee race, approximately 10% of arrests in the data. These rates are calculating using all arrests that are observable for officers over the entire sample period and are not restricted to arrests resulting from dispatched calls for service. The test sample for this data is 7/2015-10/2017.

Jail Data Officer level measures also include characteristics of jail bookings and stays. Jail booking data was obtained through an Open Records Request to the Dallas County Sheriff's department. This data includes records of all jail inmates booked by DPD officers from 7/2014-3/2018. The jail records include arrestee names, gender, race, booking date, as well as information on mode and reason for release from jail. This data also includes information on the officer(s) that booked an individual in the county jail. These officer identifiers are used to construct the total jail admissions for each officer during this period. The data is also used to calculate the proportion of jail admissions where the reason for release was that an individual was detained for a Class C Misdemeanor only and no other charge was upheld. The test sample for this data is 7/2015-10/2017.

Court Data Officer level measures also include County Criminal Court records for offenses other than Class C Misdemeanors (these citation charges are processed through an alternative Municipal Court System). This data is available for the period of 7/2014-3/2018. The court records include charge level, offense date, defendant name and case outcome information. This data is matched to arrestee-level data. The arrestee data is drawn from the combined Arrest Outcome dataset over the same period as the court data, with the restrictions that arrests need to include charge type information and Class C Misdemeanors are excluded. Individual arrests are matched when the first two letters of their first name and first three letters of their last name agree, there are no more than 2 character differences between the first names across the data sets and no more than 2 character differences between the last names across the data sets. The match rate in this sample is $> 85\%$. Conviction and dismissal outcomes are calculated at the incident, so an incident can include a conviction and dismissal if there are multiple charges. Conviction and dismissal rates are calculated among arrests that are able to be linked to the court data. The test sample for this data is 7/2015-10/2017.

Use of Force Data Data on non-shooting officer use-of-force data comes from "Response to Resistance" datasets available on the DPD website. The data covers incidents through 2016. This data includes records for all instances where officers use force, the officer badge number, demographic characteristics of the officer and civilian, whether there was a civilian or officer injury and whether the civilian resisted or was armed. The data also indicates whether the interaction occurred during the course of an arrest. I use this data set to construct a measure use of force, defined as an incident where physical force was used or a civilian was injured, the officer was not injured, the civilian did not resist and was not armed. The numerator of these rates is a count of use-of-force interactions that occur during arrests for each officer. The denominator for each officer is a count of all arrests from the combined Arrest Outcome data described above. The test sample for this data is 7/2015-12/2016.