Variation in Racial Disparities in Police Use of Force*

Carl Lieberman[a]

[a]*Center for Economic Studies, U.S. Census Bureau, 4600 Silver Hill Road Suitland, MD, USA*

Abstract

I examine how racial disparities in police use of force vary using new data covering every municipal police department in New Jersey. Along the intensive margin of force severity, I find disparities that disfavor Black subjects and are larger at higher force levels, even after adjusting for incident-level factors and using new techniques to address selection bias. I then extend empirical Bayes methods to estimate department-specific racial disparities and observe significant differences across and within these hundreds of departments. My findings suggest that ignoring heterogeneity in police use of force misrepresents the problem and masks the existence of both departments with very large disparities and those without apparent disparities against Black civilians, but the variation even within departments may make identifying and treating inequitable policing difficult.

Keywords: policing, police use of force, race, racial disparities
JEL: J15, K42

Racial inequities in police use of force are among the most significant social issues in the United States, with numerous calls for change but little widespread reform. Yet policing is not monolithic: it is a public good typically provided at the local level, and the extent to which there are racial inequities, or the solutions to those issues, may vary across departments. Moreover, these disparities could change as we consider different levels of force, from restraining the subject to discharging a firearm. One municipality's department might use excessive force against Black subjects concentrated at the highest levels of force, while a neighboring department might have disparities only at lower levels of force, and another might not have any apparent racial differences. To what extent is this true, and what does this mean for identifying areas for policing reform?

I examine fundamental aspects of how racial disparities in police use of force vary to better understand these problems. First, I explore aggregated data to estimate key parameters of race and policing in my setting: how large are overall racial disparities in police use of force, and how do they change along the spectrum of force severity? I then consider variation in these disparities at the department level: how do these racial differences vary across and within departments and their localities? I answer these questions using novel administrative data from New Jersey. These data offer an unprecedented level of breadth and depth for police use of force data in a unified setting, containing every use of force report by every officer in every municipal police department and the state police over the five-year span from 2012 through 2016. Because the data cover the entire state,

*Any opinions and conclusions expressed herein are those of the author and do not reflect the views of the U.S. Census Bureau.

Email address: `carl.lieberman@census.gov` (Carl Lieberman)
URL: `https://carl.ac` (Carl Lieberman)

they feature substantial variation in locations across hundreds of departments and municipalities. This presents a significant advantage over data used in prior research that are often based on a handful of large urban departments willing (or compelled by the courts) and able to share their use of force data, enabling not only more credible and representative estimates of racial disparities, but also new measures of how they vary.

For the initial question of how overall racial disparities vary over the force spectrum, I focus on the intensive margin of force: conditional on force being used and adjusting for event-level observables such as incident type and subject actions, do people of different races have different levels of force used against them? This parameter is relevant both in the empirical literature and in public policy, reflective of the higher levels of force experienced by Black civilians in my and others' data analyses and in numerous high-profile incidents of police violence recently and historically. To estimate these empirically challenging parameters, I use several different methods, including an extension of the "veil of darkness" test from Grogger and Ridgeway (2006) and identifying a set of incidents where the subject's race was unlikely to factor into the extensive margin process of whether force was used at all, limiting selection bias. I find significant disparities in police use of force across races that are larger at more severe levels of force. I also argue that department-level variation plays a large role, which contrasts with prior studies that focus on individual officers, but may better be able to explain the endemic problems with some departments that have resulted in the nearly two dozen departmental consent decrees currently ongoing from the Department of Justice.

Next, I explore how racial disparities vary across and within departments with new empirical Bayes techniques for group-specific differences. Focusing on overall disparities masks significant variation across departments that can dwarf the full-sample estimates: a minority of departments does not have estimated racial disparities against Black subjects for a given force level, but there are also long tails of departments where they face disparities significantly larger than the overall estimates. I also find that while the presence of a disparity in force usage against Black subjects at one level of force severity makes a disparity at another force level much more likely, many departments have disparities at only some levels of force. In supplemental analyses, I explore the local and municipal characteristics that may be predictive of a department having a large disparity, finding some evidence of a role for municipal factors, but results are not conclusive, and the variation within departments complicates efforts to identify (and treat) inequitable policing.

Although there is little work specifically about variation in racial differences in police use of force, researchers have made progress on questions of race and criminal justice using new data and empirical strategies.[1] Nix et al. (2017) and Weisburst (2019), for example, use new data from *The Washington Post* and the Dallas Police Department, respectively, both finding racial disparities against Black civilians in different aspects of police use of force. Fryer (2019) estimates overall disparities by combining several different data sources, including New York City's Stop, Question, and Frisk program and officer-involved shootings from 16 police departments. He finds racial disparities in police use of force for Black subjects relative to White subjects to be consistent

[1]Related topics besides use of force include police stops and searches (Knowles, Persico, and Todd, 2001; Anwar and Fang, 2006; Persico and Todd, 2006; Antonovics and Knight, 2009; Coviello and Persico, 2015; West, 2018), speeding tickets (Goncalves and Mello, 2021), police reform after racially unjust incidents (Heaton, 2010; Devi and Fryer, 2020; Luh, 2020), whether subjects are formally booked (Raphael and Rozo, 2019), and court decisions (Rehavi and Starr, 2014; Arnold, Dobbie, and Yang, 2018; Bielen, Marneffe, and Mocan, 2019; Tuttle, 2021).

over the nonlethal force spectrum with odds ratios of around 1.2; he also controversially finds no disparities for shootings.[2] Ross (2015) compares county-level police shootings with local racial populations and finds that risk ratios are greater in "larger metropolitan counties with low median incomes and a sizable portion of Black residents, especially when there is high financial inequality." Nicholson-Crotty, Nicholson-Crotty, and Fernandez (2017), Hoekstra and Sloan (2020), and Ba et al. (2021) all suggest a positive impact from an increase in the proportion of Black officers. Cunningham, Feir, and Gillezeau (2021) cite the expansion of collective bargaining rights for officers as a contributor to racial gaps in police killings. And in labor economics, my findings of department-level disparities are paralleled by the result in Kline, Rose, and Walters (2022) of company-specific racial disparities in job applicant callback rates persistent across space and time.

In identifying these variations in racial disparities, I move forward our understanding of race and policing. Beyond simply observing heterogeneity in police use of force, the variation present in department-level disparities underscores the difficulties in remedying those differences and how there may be gains from more targeted interventions. These issues are important to understand on their own, but they also contribute to a much needed holistic understanding of race and criminal justice. Consider, for example, how if officers could use less force overall without compromising policing quality, it would reduce the number of subjects affected by disparities in force intensity, or how greater levels of force against Black civilians could affect the overall efficiency of a police department given evidence of decreased public cooperation and engagement after high-profile incidents of police violence (Ang et al., 2021). Identifying and addressing racial inequities in police use of force is a vital step, but it is one on a much longer journey.

1. Institutional Background and Data

1.1. Use of Force and Reporting in New Jersey

Per the New Jersey Attorney General's use of force policy from the sample period, officers in New Jersey are instructed to use force to facilitate an arrest or "other law enforcement directive," and the level of force used "should be only that reasonably necessary" after officers "exhaust all other reasonable means before resorting to the use of force" (McCarthy and Nelson, 2019; New Jersey Attorney General, 2000). In 2001, the New Jersey Attorney General's office began to require that police officers document all incidents in which they use force. Appendix Figure A.1 shows a model form for these force reports. Plans for a centralized system for collection, oversight, and analysis of these reports by the state did not materialize, and most of the reports ended up in storage, unused and inaccessible by the public (NJ Advance Media, 2019).[3]

The force types police in New Jersey are authorized to use can be ordered along a spectrum of severity. The lowest level is compliance holds, maneuvers like arm bars and wrist locks that use pressure points to restrain the subject. Next are unenhanced types of force: takedowns (forcing a subject to the ground, like tackles), slapping, punching, and kicking. Then there are enhanced

[2] Knox, Lowe, and Mummolo (2020) argue that the estimates from Fryer (2019) understate the causal effect of race by not accounting for racial differences in police-civilian interactions prior to the encounter, such as patrolling habits and the decision of whether to stop a subject. See also comments by Durlauf and Heckman (2020) and the response by Fryer (2020).

[3] In the wake of the murder of George Floyd by police officer Derek Chauvin in Minneapolis, the New Jersey Attorney General developed a statewide use of force portal that launched in April 2021.

force outcomes like pepper spray and other chemical agents, baton strikes, canines, and stun guns (Tasers).[4] Finally, as a subset of enhanced force, officers may use deadly force by discharging firearms. Police very rarely employ nonlethal firearm ammunition like beanbag rounds in settings such as riots; I drop these incidents, as these situations are not representative of typical police-civilian interactions. The threat of force without its use ("constructive authority"), like drawing a firearm, is permitted, but warning shots are prohibited.

1.2. Data

This project uses all known force reports from every municipal police department in New Jersey and the New Jersey State Police from 2012 through 2016. ProPublica, a nonprofit newsroom, and NJ Advance Media (NJAM), a news provider that operates NJ.com and *The Star-Ledger* newspaper, obtained the reports through more than 500 public records requests and several legal threats. Among other variables, the reports record the time, date, and location of each incident, its type (such as a crime in progress or traffic stop), the officer(s) involved and their demographics, the subject(s) involved, their demographics, and their actions that led to force being used, the type(s) of force an officer used against the subject(s), and whether an officer or subject was injured or killed.

The final dataset from ProPublica, NJAM, and their partner data entry firm requires additional processing before use in analyses. For example, some types of force used by police consist of irregular descriptions such as "grabbed rock out of her hand." Appendix A documents in detail how I process the data into a consistent format. I structure the data so that each observation represents one subject who has force used against them by one officer in one incident. For cases where multiple officers use force against a single subject, I keep the officer who uses the greatest level of force, choosing randomly in the event of a tie.[5] I remove 44 subjects whose races recorded on the force reports do not fall within the categories of White, Black, Asian/Pacific Islander, or Hispanic,[6] such as people marked as "mixed." After cleaning, there are 39,321 incidents that I use in my primary analysis sample, with the exact number varying slightly depending on the model used due to certain variables' missingness.

I supplement the force reports with data on New Jersey's police departments and municipalities. Characteristics of the police departments themselves, such as number of officers, come from ProPublica and NJAM, with some additional data in supplemental exercises coming from the 2016 Law Enforcement Management and Administrative Statistics (LEMAS) (Bureau of Justice Statistics, 2016). Local variables come from the 2010 decennial census, 2012-2016 American Commu-

[4]Stun guns are mostly absent from the data due to state regulations against their use; although the law changed in 2016 to permit stun guns, adoption was slow, and they are too infrequent in the data to know whether an officer in a given incident may have been carrying one. Canines are also uncommon, as not all departments have canine units.

[5]There are several justifiable ways to structure the data. I choose this format because it is the most natural for my empirical strategy in which the outcome of each incident is based on the greatest level of force used. When multiple officers use force, it is then sensible to keep the officer who used the greatest level of force, especially because other officers may only be able to use lower levels of force due to the officer using a more severe type (for example, after Officer A pepper sprays a subject, Officer B is able to place them in a compliance hold).

[6]Due to data limitations, I treat Hispanic status as a distinct racial category; one cannot be Hispanic and another race in the data. Although some departments' force reports do distinguish between a subject's race and Hispanic status, many do not, and the cleaned version of the race variable in the dataset made available to researchers treats Hispanic status as a race. Officers use their own judgment when recording a subject's race.

nity Survey (ACS) five-year estimates (Ruggles et al., 2020), the FBI's Uniform Crime Reporting (UCR) program, the New Jersey Division of Elections, and Harvard's Project Implicit.

1.3. Summary Statistics and Stylized Facts

Table 1 presents summary statistics for the cleaned force reports. Force severity is inversely correlated with frequency: half of all incidents involve only compliance holds, while officers discharge firearms in under half a percent of incidents. Subject actions are similar, with resisting an officer (which I also use as a catchall for non-missing actions that do not fit elsewhere) and physical threats/attacks combining for 97% of the data. Following the model force report from the New Jersey Attorney General, I classify incidents as crimes in progress, domestic disputes, other disputes, suspicious persons, traffic stops, and other incidents. An incident may have multiple types, except for the "other" category, which I reserve for incidents that do not belong to any of the other categories mentioned above. While the modal incident is a crime in progress (27% of the sample) and traffic stops are only 9% of the sample, in the 2015 Police-Public Contact Survey, among US residents whose most recent police contact was police-initiated, more than 85% of these incidents were traffic stops; only 6.7% were street stops (Davis, Whyde, and Langton, 2018).

Subject demographics are not representative of New Jersey's population. 48% of the subjects in the sample are White, 41% are Black, 10% are Hispanic, and 1% are Asian/Pacific Islander. In the ACS estimates for 2012-2016, the state's population was 57% non-Hispanic White, 13% non-Hispanic Black, 9% non-Hispanic Asian/Pacific Islander, an 19% Hispanic of any race. While the differing treatments of Hispanic status between the force reports and ACS makes determining the exact extent to which a group is over- or under-represented impossible, Black subjects make up a large enough portion of the force reports that they must be overrepresented relative to their population proportion.

Appendix Table A.1 contains summary statistics for the municipal police departments in New Jersey and the State Police. Department sizes vary greatly, with a median of 23 full-time employees and a maximum over 1,000. Racial diversity is low, and the median department reports that more than 90% of its officers are White.[7] This is closely related to population demographics, and the correlation between the percentage of officers in a department who are Black and the corresponding percentage of residents ages 18-65 in that municipality is 0.68. Many departments in New Jersey have residency requirements that officers live in the municipalities they work, which may affect officer diversity. Departments make arrests far more often than they use force, with force incidents totaling about 3% of arrests overall, for Black subjects only, and for all other subjects.

Appendix Table A.2 presents summary statistics for the 454 municipalities with their own police departments that are present in the cleaned data.[8] Municipalities are heterogeneous, with populations ranging from the hundreds to the hundreds of thousands, White population shares from 2% to 99%, and violent crime rates from 0.03 per 1,000 residents to 25.7 per 1,000. County-level Black-White implicit association test (IAT) D-scores from Project Implicit, where more positive values indicate stronger implicit biases against Black people, also show large variation: the standard deviation of 0.05 is approximately equal to the difference between the median and 70th percentile counties at the national level.

[7] Due to inconsistencies in reporting, not all departments' racial breakdowns sum to 100%.

[8] A small number of municipal departments also cover a neighboring municipality that lacks its own department. These neighbors are very small, with a median 2020 Census population of 1,003.

Appendix Figure A.2 looks at how often departments use force overall and relative to arrests and how these relate to race. Across all departments, the total number of incidents per arrest resembles a log-normal distribution, slightly left-skewed, with large variation. Departments vary greatly in their force frequencies even after normalizing by arrests. The 90th percentile department has 5.3 times as many force incidents per arrest as the 10th percentile department, and the maximum department uses force 88.0 times more often per arrest than the minimum department. Breaking down the total force incidents per arrest into Black force incidents per Black arrest (and the non-Black equivalents) shows that the distributions are fairly similar across races. But the Black distribution has greater variance, and both the maximum and minimum force incidents per arrest values are greater than they are in the non-Black distribution.

The proportions of force subjects, arrestees, and the local population ages 18-65 that are Black are closely linked, plotted in Appendix Figure A.3. Areas with the lowest percentages of Black residents not only have disproportionately high shares of Black force subjects and arrestees, but also the greatest marginal increase in those outcome shares as the Black population increases. Black arrest and force incident proportions rise near a 1:1 rate. The concave shape of the force subjects-population relationship persists in a regression of each department's proportion of force subjects who are Black on a vector of local characteristics: a quadratic of their local population ages 18-65 that is Black, log population, log population density, and violent crime rate (see Appendix Table A.3). Other models with the same regressors find that departments' total number of force incidents is increasing in population size, density, and violent crime rates, with another positive concave relationship with the 18-65 Black population, and that force incidents per arrest are increasing in violent crime rates, with a small but statistically significant positive correlation with population density.

2. Empirical Strategy

2.1. Model Specifications

2.1.1. Overall Disparities

My primary econometric specification estimates the following equation via ordinary least squares (OLS):

$$Force_{iopt} = \beta \cdot \text{Black}_i + X'_{iopt}\gamma + \psi_p + \nu_t + \varepsilon_{iopt} \tag{1}$$

where the subscripts i, o, p, and t denote the incident-subject pair, officer, department, and year, respectively. *Force* is a binary measure of whether the level of the force used in an incident was at least a certain severity: compliance holds, unenhanced force, enhanced force, and firing a weapon (a subset of enhanced force) For example, officers strike a subject with a baton, the outcome for at least compliance holds, at least unenhanced force, and enhanced force would be one, but it would be zero for firing weapons. Because compliance holds are the lowest level of force, the at least compliance hold measure is always one, and I omit it as an explicit outcome. This force parameterization is an intuitive way to interpret outcomes, with alternative methods like the maximum force used or each force level individually incompletely describing events or resulting in estimates without clear meaning (Fryer, 2019). The coefficient of interest is β, the difference in the observed probabilities of Black subjects having more severe types of force used against them conditional on force being used at all relative to people of other races and after adjusting for incident-level factors. X is a vector of incident characteristics including time, type of incident, officer rank, subject behaviors, whether they were marked as "emotionally disturbed," subject sex, and a quadratic of the subject's age. I include indicators for each unique combination of incident types rather than each individual type. Department fixed effects ψ_p capture time-invariant aspects of each department's propensity to use higher levels of force, like policies, training, and municipal characteristics. Including these fixed effects allows for separating these impacts from the estimand of interest of residual racial disparities, eliminating channels such as sorting between race and a department's race-neutral force propensities. Time fixed effects ν_t adjust for year-month specific changes in overall force usage to capture seasonality in crime. I also estimate an additional series of models that use officer fixed effects and clustering instead of department. These serve a similar purpose but can adjust at a finer level for factors such as one officer having a dangerous beat while another works primarily on traffic enforcement, which impacts force usage. Further, comparing these results to the ones using department fixed effects allows for better understanding the roles of the officer and the department.

For robustness, I also estimate conditional logit models of the form

$$\ln\left(\frac{Pr(Force_{iopt}=1)}{1-Pr(Force_{iopt}=1)}\right) = \beta \cdot \text{Subject Black}_i + X'_{iopt}\gamma + \psi_p + \nu_t + \varepsilon_{iopt} \tag{2}$$

stratified by department. To prevent separation,[9] I use year fixed effects instead of year-month, use incident type indicators for each individual incident type, change the officer type indicators to a dummy for having a superior rank, and do not estimate models with officer fixed effects. Conditional logit has the advantages of bounding probabilities between zero and one addressing issues of inconsistency of logit with numerous fixed effects. However, at greater levels of force, more departments will have only null outcomes, and they must be dropped with logit-based methods. I also run two ordered logit models to examine the entire force spectrum of force severity simultaneously. One model uses the same regressors as Equation 1 to examine the full spectrum of force at once, and the other augments the force reports with aggregrated municipal arrest records from the FBI (Kaplan, 2021), treating arrests as an outcome lower than force. This precludes the use of detailed incident microdata but allows for incorporating a measure of the extensive margin.

2.1.2. Department-Specific Disparities

The simplest approach to estimating departmental racial disparities would be to make a slight modification to Equation 1 to run regressions of the form

$$\text{Force}_{iopt} = \beta_p \cdot \text{Black}_i \times \text{Department}_p + X'_{iopt}\gamma + \psi_p + \nu_t + \varepsilon_{iopt} \tag{3}$$

with the interest being in the distribution of β_p. Although these estimates may be unbiased and consistent, in general, OLS will tend to generate the most extreme point estimates for departments with the fewest observations. I propose an alternative approach that can better handle estimates

[9]Albert and Anderson (1984) define separation in logistic regression as the situation when "there is a vector α that correctly allocates all observations to their group." As an example, imagine an officer with several force incidents, all of which involved discharging a firearm, so the outcome is always one for their incidents. In an OLS regression, the officer fixed effect is estimable and interpretable. But with logistic regression, there is a problem because the maximum likelihood estimate does not exist: the larger the officer fixed effect, the larger the likelihood, and so the fixed effect diverges to infinity and the model breaks.

from all departments.

I modify empirical Bayes estimators, sometimes used by economists for estimating teacher value-added (see, for example, Kane and Staiger, 2008, and Chetty, Friedman, and Rockoff, 2014), to estimate group-specific differences for each department. Empirical Bayes estimators use the overall distribution of estimates to inform each individual point estimate, with a prior distribution calibrated on the observed data instead of being chosen as with pure Bayesian methods.

As is standard practice, I start by fitting a normal distribution for the prior and apply Bayesian updating to obtain a posterior distribution for each department's point estimate. Less reliable estimates, such as those from departments with few observations, are shifted more towards the population mean, resulting in a "shrinkage" estimator. After applying the updating procedure, I record the centers of the posterior distributions as the estimates of β_p.

For each level of force, I begin by running the following "pooled" regression. Note that there is only one overall coefficient for the subject being Black, hence all departments are pooled together.

$$\text{Force}_{iopt} = \beta_0 \cdot \text{Black}_i + X'_{iopt}\gamma + \psi_p + \nu_t + u_{iopt} \tag{4}$$

I use $\beta_p \sim N(\hat{\beta}_0, \sigma^{2_p})$ as the prior, where $\sigma_p^2 \equiv \text{Var}(u^B_{iopt} - \varepsilon^B_{iopt})$, the variance of the difference of residuals between Equations 3 and 4 using only Black observations rather than the full sample. I then take the estimated racial disparities β_p from Equation 3 and compute the empirical Bayes estimates, which simplify to weighted averages of the prior mean and the department estimate.

$$\hat{\beta}_{p,EB} = w_p \cdot \hat{\beta}_p + (1 - w_p) \cdot \hat{\beta}_0 \tag{5}$$

where

$$w_p \equiv \frac{\sigma_p^2}{\sigma_p^2 + \text{Var}(\varepsilon_p^B)/n_p^B}$$

are the departmental empirical Bayes weights, ε_p^B is the residual for an observation with a Black subject in department p (residuals for other subjects are omitted), and n_p^B is the number of observations with Black subjects for department p. Due to small sample issues for certain departments, I winsorize the estimated disparities at the 1st and 99th percentiles.[10]

2.2. Identification and Limitations

Every incident undergoes two treatments prior to the officer's decision of what level of force to use: whether to engage with the subject, such as pulling over a vehicle or making a stop on the street, and whether to use force. Each of these is plausibly affected by race, and analyses in other settings have found evidence that they are (Gelman, Fagan, and Kiss, 2007; Fryer, 2019). Because my data are observed after these decisions, selection bias in who is in the sample, the population with force used against them at all, could make interpretation of these disparities difficult. Intensive margin racial disparities in force intensity are consistent with positive, negative, or null extensive margin racial differences, and such differences could arise from many possible

[10] Although empirical Bayes estimates are designed to shrink less reliable point estimates to the center of the distribution, in extremely fine samples like the smallest departments, the model may fit the data almost perfectly, resulting in minimal shrinkage. Consider a department that uses force against only two similarly aged Black subjects, as part of the same incident and using identical levels of force.

causes with different behavioral interpretations and policy implications.[11] And the direction of subsequent bias from extensive margin racial differences can be ambiguous. Would a greater exposure rate of Black civilians to police (like from police discrimination, overpolicing, or bias in 911 calls) cause less threatening Black subjects to appear in the data, or would police be able to intervene in more violent crimes as a result?

The easiest way to understand how selection bias could manifest is akin to an omitted variables framework. Imagine that force reports contained a perfectly accurate measure of the threat level/dangerousness/criminality of a subject, denoted $Threat$. Naturally, $Threat$ is positively correlated with higher levels of force, so the direction of omitted variable bias from excluding $Threat$ matches the sign of the correlation between the subject being Black and $Threat$: factors that increase/decrease $\text{cor}(Black, Threat)$ would shift intensive margin disparity estimates upwards/downwards. Although the models do attempt to control for $Threat$ by adjusting for incident and subject characteristics and actions, these shifts can occur even with perfect adjustments. Focusing on the linear probability models for simplicity, the coefficient on the subject being Black estimates the average probability that a Black subject experiences the higher level of force, all else equal. But racial differences in the extensive margin selection process could, for example, result in low-threat Black subjects who are never on the margin for having higher levels of force used against them, shifting the intensive margin coefficient for the subject being Black downward despite no changes in the intensive margin process.

Extensive margin differences are also important to consider on their own as a component of equitable policing. Many of the practices that would result in less-threatening Black subjects having force used against them are consistent with bias and/or overpolicing relative to people of other races, while those that increase the Black-threat correlation might suggest an ability to identify and intervene in more severe incidents, which could be beneficial if applied generally, or they could be a result of police antagonizing and provoking subjects.

In my first approach to understanding and addressing possible biases from selection into the sample, I build on the veil of darkness test from Grogger and Ridgeway (2006) and Horrace and Rohlin (2016), which compares the ratio of Black to other motorists stopped during daylight (when driver race is more likely to be observable by the officer before the stop) and in darkness within a fixed set of hours that are both light and dark over the year. Absent strong conditions, it is a qualitative test of whether there is discrimination against Black drivers in traffic stops. In addition to darkness, I consider "always force" incidents where the subject at least physically threatened or attacked an officer or another individual and the outside option of not using force at all was less viable. Formally, this allows me to "veil" the extensive margin force decision and benefit from the veil test's milder assumptions and its ability to account for the unknown at-risk population and possible nonreporting.[12] I outline this approach in greater detail in Appendix B. This exercise

[11] This is related to the more general idea of differences in the marginal subject across races discussed, for example, by Becker (1957).

[12] NJAM has attempted to detect and fix missing or incorrect reports through a crowdsourcing effort (McCarthy, 2019). This has uncovered minor discrepancies in some reports, often around officer names, and the only missing reports uncovered are 70 from Jersey City (prior to cleaning, there are more than 1,100 force reports from Jersey City). Reporting by NJAM's NJ.com and a discussion with an active New Jersey police officer also indicate that there is some disagreement about when officers are required to fill out force reports. In the main analyses, because the coefficient of interest is the racial disparity after adjusting for incident characteristics, a necessary condition for

provides evidence on the extent of *extensive* margin racial disparities within these traffic stops, of interest as its own result and to help understand the direction of possible selection bias.

While the extended veil test confers a high degree of internal validity, it is narrow in scope by focusing on a subset of vehicle stops. To examine a broader set of incidents while attempting to limit the role of the subject's race prior to the intensive margin of force intensity, I repeat my main analysis of overall racial disparities the intersection of "always force" incidents and a new set of "always stop" incidents. These "always stop" incidents are cases where the subject's race is less likely to factor into the officer's initial decision to engage with the subject at all, either because race is difficult to observe or the incident's severity is sufficiently high that officers would engage with the subject under any circumstances: crimes in progress, disputes, and traffic stops in the dark. While these incidents are "veiled" in a sense, this is a fundamentally different test than the extended veil of darkness and is akin to"identification at infinity" from Heckman (1990). Differences between the full sample disparities with the subsetted disparities would suggest some degree of extensive margin racial differences impacting results.

Despite the reduced number of observations (8,050), this subsetted sample retains a high level of representativeness.[13] All racial/ethnic and sex proportions are within one percentage point of the full sample values in Table 1, and the mean age difference is less than half a year. 93% of departments are present in both datasets and missingness is similar across different parts of the state, with dropped municipalities tending to be smaller, Whiter, and richer; the corresponding departments are also smaller and Whiter than the full sample's. The largest difference is that dropped municipalities tend to have very low violent crime rates, consistent with the subsample consisting of more severe incidents. For similar reasons, 37% of officers are present in both samples, skewing towards those who use force more often.

My final approach to possible selection bias is adapted from Goncalves and Mello (2021) and uses the logic of the Heckman correction, formally derived in Appendix C. Under the Heckman correction model, by estimating each department's race-specific extensive margin force propensity (such as from treating arrests as the universe of incidents where force might be used, an assumption discussed in Appendix C), we can identify the expectation of any such selection bias by adding the Mills ratio of these propensities to the regressions. An insignificant coefficient on the Mills ratio and lack of change in the estimated racial disparities or other coefficients would then be evidence supporting a lack of selection bias in intensive margin analyses.

these to bias my estimates is that any errors in the data such as missing reports are correlated with the subject's race. If officers systematically misreport Black subjects as posing greater threats than they actually do, this would cause me to underestimate racial disparities. Incentives to misreport are greatly diminished by the lack of central oversight during the study period. Force reports existed mostly as physical copies, many only at the departments themselves despite guidelines that all force should be reported to county prosecutors, making external monitoring difficult (Nelson, 2019; McCarthy and Stirling, 2019).

[13] For summary statistics on the subsetted sample, see Appendix Tables A.4 for its force reports, A.5 for its municipalities, and A.7 for its departments. For summary statistics on the dropped groups only, see Appendix Tables A.6 for municipalities and A.8 for departments.

3. Results

3.1. Overall Racial Disparities

3.1.1. Estimation and Robustness

Figure 1 plots the racial disparities from the full model in Equation 1 for Black subjects on top of outcome means, with corresponding regression results in Table 2. In the department-based models, the disparities estimated for Black individuals in the at least unenhanced, enhanced, and fired weapons outcomes are 1.9, 3.3, and 0.1 percentage points, respectively, with all except the firearms level statistically significant at the 1% level. While the positive disparities for firing weapons have economically significant point estimates given the rarity of these events, that rarity greatly increases uncertainty, and the confidence intervals are wide.[14] Relative to the frequencies with which each force type is used, disparities are greater at higher levels of force: the department fixed effects-based disparities for Black subjects are 4%, 23%, and 13%. Results are similar when using officer fixed effects and clustering, with larger standard errors; adjusting for individual officers instead of the department does not materially change estimates of racial disparities. Estimates with a full set of race indicators in Appendix Figure A.5 and Table A.10 show that relative to White subjects, Hispanics and Asians/Pacific Islanders do not have significantly different outcomes, except that Asians/Pacific Islanders have a relatively large *negative* disparity for the fired weapons outcome that is highly significant with the more precise standard errors in the department-based model. For a discussion of individual departments' partial leverage (similar to weights from Goodman-Bacon, 2021), influence, and identifying variation for these disparities, see Appendix D.

Results are robust to functional form choice. The odds ratios from the conditional logit models in Equation 2 presented in Appendix Figure A.6 and Table A.11 are qualitatively similar to the OLS results. Point estimates are positive and small at the lower levels of force and larger at higher levels, with the odds ratios for the at least unenhanced, enhanced, and fired weapons outcomes being 1.05, 1.29, and 1.09, respectively, and the former two are significant at the 1% level. The ordered logit model in the first column Appendix Table A.12 examining the entire ordered force spectrum simultaneously is also similar, with highly significant odds ratios on the subject being Black of 1.15. And results in the second column, which incorporate arrests as the lowest outcome but cannot adjust for most incident-level factors,[15] again show a large (1.30) and highly significant disparity.

Turning to the checks for selection bias, the extended veil test results in Table 3 suggest an extensive margin disparity against Black subjects across a variety of specifications and incident subsets. Including all intertwilight traffic stops may improve power, but examining only evening incidents results may strengthen the test's constant risk ratio assumptions, especially given racial differences in when people work (Hamermesh, 1996). Despite limits in power from the number of intertwilight traffic stops involving force, all odds ratios are greater than one and significant at the 10% level, and the model that uses all intertwilight incidents and adds the racial population control is significant at the 5% level. Figure 2 and Appendix Table A.13 contain the results from regressions on the "always stop" and "always force" incidents. Point estimates are larger in magnitude than their full sample analogs, but outcome means are naturally higher here than in the full

[14] Contrasting Fryer (2019), I find disparities against Black subjects for shootings even in the raw data without controls (see Appendix Figure A.4 and Table A.9).

[15] I discuss the assumption that all force incidents involve arrests in Appendix C.

sample, as these incidents tend to be more severe than the omitted ones, prompting higher levels of force. Relative to the new outcome means, the results are similar to and not statistically significantly different from the full sample ones (6%, 26%, and 14% for the three outcome levels), and this similarity suggests a lack of large selection bias in the full sample estimates. Finally, in all three Heckman correction-style models in Table 2, the coefficient on the Mills ratio is statistically insignificant at the 5% level, with two of the three also insignificant at the 10% level (the p-value for the fired weapons model is 0.08). Moreover, the racial disparities and other coefficients are effectively unchanged from the main three regressions, suggesting minimal impact of any extensive margin racial differences.

Two of these selection bias checks, the subsetting and Heckman correction exercises, are consistent with a lack of bias from differential selection into the sample. The extended veil test suggests that Black subjects are more likely to have force used against them in traffic stops. When thinking about how that disparity could affect intensive margin disparities, we can return to the question of whether these additional force incidents are bringing in Black subjects that increase or decrease the correlation between race and threat. Because Black subjects make up a greater proportion of force subjects when it is more likely that the subject's race was observable to police and the officer had discretion in whether or not to use force, with the latter condition in particular associated with less severe incidents, we may expect that this extensive margin disparity would result in more Black subjects with low threat levels. This could decrease the correlation between the subject being Black and the threat posed by a subject, which would shift intensive margin disparities downwards, an underestimation.

3.1.2. Mechanisms

To better understand the types of incidents driving racial disparity estimates, Figure 3 plots actual force usage against predicted force by race. The x-axis shows predicted force probabilities from the linear probability model in Equation 1 without the Black subject indicator, putting incidents into a race-neutral context, and the y-axis shows the actual and predicted proportions of incidents involving force at least that severe. These graphs allow us to see the types of incidents where Black subjects are actually experiencing greater levels of force. I omit the fired weapons outcome, as the relative rarity of police shootings results in the predicted probabilities effectively having a mass point at 0. Without racial disparities in force usage, we would expect the curves to be identical 45-degree lines. In actuality, the Black curve is slightly higher everywhere, with the largest disparities when the probability of using at least unenhanced force is about 0.25 to 0.5. Because most incidents have a probability of at least unenhanced force near 0.5, this suggests that the *overall* disparities I estimate at this force level are coming from incidents where there is an approximately 50% chance that an officer uses greater levels of force, incidents in which officers may have the most discretion in their choice, incidents in which officers may have the most discretion in their choice For enhanced force, there is a statistically significant local maximum disparity when the probability of enhanced force is near 0.25, near the modal predicted probability. Like with unenhanced force, these incidents drive the overall disparity estimates and may allow for more officer freedom in force level choice.

We may ask whether these disparities are the result of taste-based or statistical discrimination, as the policy implications for each may be different. Note that racially discriminatory policing for any reason is illegal in New Jersey and should not be considered justified regardless of motivation. Statistical discrimination may manifest due to officers' belief or stereotype that, all else equal, Black subjects pose greater threats than subjects of other races, prompting higher levels of force (these beliefs need not be accurate; see Bohren et al., 2020). Under the assumption that male officers should not believe that they need this extra force to respond to female Black subjects, interacting race with the female indicator for incidents with male officers can differentiate between these channels. A negative coefficient on the interaction between the subject being Black and female would suggest statistical discrimination, while a null coefficient would be more consistent with taste-based discrimination. Table 4 shows that coefficients from this augmented model on the subject being Black, the subject being female, and their interaction are positive, negative, and approximately zero, respectively, for both the at least unenhanced and enhanced models, supporting the taste-based discrimination hypothesis (I do not include the fired weapons outcome, as only five women are shot in my sample).

Table 5 focuses on harm to officers. It presents results from a version of Equation 1 that moves the officer harmed[16] indicator to the left-hand side and puts indicators for the maximum level of force used, also interacted with the subject being Black, on the right-hand side. Officers are significantly more likely to report being harmed in incidents where the maximum level of force used was greater than compliance holds. Although these higher levels of force are intended to protect officers and others, we cannot observe the timing of the subject's actions and the officers' use of force, and it is likely that this relationship exists at least partially because officers use higher levels of force after the subject harms someone. Incidents with Black subjects are a statistically significant 1.8 percentage points more likely to have an officer report being harmed, and interactions with subject race and force level are all insignificant at the 5% level and inconsistent in sign. That officers report being harmed more often in incidents involving Black subjects, even after adjusting for the subject's actions, incident type, etc., is difficult to reconcile with officers using greater levels of force against them. One explanation could be officers misreporting or overreporting harm to themselves to justify their greater force usage against Black subjects; this is likely much easier to do than underreporting or not reporting their own force usage. This would suggest that estimated disparities against Black subjects are underestimated, with some of this greater force inaccurately being attributed to harm to officers rather than to the subject's race. It is unlikely that this result arises from officers antagonizing or provoking Black subjects, as that could be captured by the regressors for the subject's actions in an incident. Regardless of explanation, these results suggest that greater levels of force against Black subjects do not protect officers.

What is the relationship between the subject's race, the number of officers using force, and the intensity of that force? Table A.14 shows estimates from an altered Equation 1 where the outcome is the number of officers who use force against a subject and the sample is limited to incidents with only a single subject. Results are highly significant, with Black subjects having about 0.08 or 0.09 additional officers using force against them after adjusting for all other incident characteristics. Having more officers use force is associated with higher levels of force, as seen in Table A.15, which shows regression coefficients from another modified version of Equation 1 where the set of regressors is augmented with the number of officers using force and interactions between those numbers and the subject being Black. There is a clear pattern in which incidents involving more

[16]This variable is based on separate fields for an officer being injured, hospitalized, or killed. 3,984 incidents, about 10%, involve an officer being harmed in some manner. 3,853 indicate an officer was injured, 673 indicate an officer was hospitalized, and one indicates an officer was killed.

officers have monotonically increasing probabilities of at least unenhanced and enhanced force usage. This is consistent with the idea that more officers will respond to more severe incidents that prompt greater levels of force, but a non-exclusive explanation is that having more officers choosing what level of force to use mechanically increases the probability that at least one of them will select a higher force level. There is no additional interaction between the number of officers using force and the subject being Black—overall, more officers use force against Black subjects, having more officers use force is associated with greater levels of force, and that higher force is not uniquely greater for Black subjects.

To explore the importance of the department, in Appendix E, I leverage spatial variation in incidents within departments to disentangle the effects of the department from geographic variation in subjects and incidents. I find that when departments use force within another department's jurisdiction, their own force patterns are a better predictor of their out-of-jurisdiction force levels than those of the home department, suggesting that different departments do use force differently and that department fixed effects do not solely capture geographic differences. Further, on a set of 1,095 force incidents by 189 manually identified officers who switch police departments, I fit Equation 1 with officer fixed effects and an additional indicator for whether the department in which an officer is working has an above-median racial disparity for the relevant level of force as estimated in Section 3.2, with results in Table 6.[17] Note that switchers may be a self-selected group and their sorting across departments is likely nonrandom. For at least unenhanced force, the interaction between a subject being Black and the department having an above-median disparity is positive, large, and statistically significant, supporting the idea that the department or municipality in which an officer works is a larger determinant of racial disparities than the identity of the officer. For enhanced force, the interaction is positive but insignificant, possibly because these rarer events have limited power, or because enhanced force may be less discretionary, leaving less room for any marginal effects of the new department on force intensity.[18] These results suggest that departments play an important role in racial disparities, which leads to the next set of results exploring to what extent departmental disparities may vary.

3.2. Variation in Racial Disparities

Figure 4 presents kernel density plots of the empirical Bayes departmental disparities, with corresponding summary statistics in Table 7. These distributions have long tails that can dwarf the overall racial disparities estimated before. For at least unenhanced force, enhanced force, and fired weapons, the standard deviations of the winsorized departmental disparities compared to the (mean departmental disparity) are very large: 0.07 (0.02), 0.05 (0.01), and 0.01 (0.00), respectively. The 95th percentile department is estimated to use at least unenhanced force 10 percentage points more often against a Black subject than it would against a comparable White subject, with a corresponding 7 percentage point gap for enhanced force. However, 18% of departments have weakly negative

[17] I do not analyze firearms here, as no switching officers use that level of force in my data. Additionally, I include year fixed effects instead of month-year due to the greatly reduced sample size.

[18] Related to the topic of the malleability of officer actions, Appendix Table A.16 adds officer experience and its interaction with the subject being Black to Equation 1. On this sample (see table notes about officer experience coverage), more experienced officers are significantly more likely to use at least unenhanced or enhanced force, but the coefficient on the interaction between experience and subject race is a precise null. Because officers' beats, assignments, and ranks are all endogenous and linked with experience, I caution against a strict interpretation of greater experiencing causing higher levels of force, but it is consistent with officers changing behaviors over time.

disparities for the at least unenhanced force outcome, 24% have them for enhanced force, and 37% for firing weapons. Although these are a minority of departments, they serve as an existence proof for departments without racial disparities against Black subjects in their use of force, at least for certain outcomes, and may be a starting point for researchers and stakeholders seeking to improve outcomes in other departments.

Summary statistics for confidence intervals for the empirical Bayes estimates based on the formulation from Morris (1983) are in the final three columns of Table 7 and are plotted in Appendix Figure A.7. Most departments have a high degree of uncertainty in their estimates, but departments in the tails are mostly highly statistically significant because the predictions tend to fit the data well for these departments, resulting in a greater level of precision and less shrinkage than departments shrunk more towards the grand mean. However, this concept of precision may be incomplete. Especially in the at least unenhanced force level, some departments with few Black subjects have extreme disparities. This can be seen in Appendix Figure A.8, which presents the relationships between a department's empirical Bayes disparities, its number of Black subjects, and its total number of incidents. While departments with few Black subjects could simply be more likely to have very large disparities, estimates from departments with few Black subjects may appear precise because the model fits their few observations well, and they may be less stable in the sense that they could change the most with the addition of new observations.

Department-specific racial differences explain a relatively high portion of variance in force outcomes. Comparing R^2 values from from the pooled regression without the department-Black interactions in Equation 4 to the unpooled regression in Equation 3, the addition of departmental disparities increases the proportion of explained variance by an average of 7%. This is much more explanatory power than even sex or incident type: removing and adding the subject's sex to the unpooled regression increases R^2 by 4% on average, and removing/adding the incident type indicators is only a 2% improvement.

Figure 5 shows how racial disparities vary within departments, presenting a series of probabilities that a department does not have a positive disparity against Black subjects. The unconditional probability that a department does not have such a disparity in the at least unenhanced force outcome is 18%. This rises to 31% if we condition on the department not having a disparity for enhanced force, 26% if conditioning on having no disparity for firearms, and 28% when conditioning on both. Similar patterns exist at each force level: the probability that a department does not have a disparity against Black subjects increases greatly when conditioning on not having a disparity at another force level. Although departments with racially equitable use of force practices at one level are much more likely not to have a disparity at another level, many departments do have disparities at some force levels despite not having them at others. This makes identification, understanding, and treatment of racial disparities more complicated.

In an effort to explore the factors linked with these departmental disparities, Appendix F discusses using random forest models to examine departmental and municipal characteristics that may predict whether a department will have an especially large disparity against Black subjects. I find that certain municipal characteristics like household income, violent crime rate, and election results are informative, but also that identifying local features that are broadly predictive of racial disparities is difficult, confounded by the variation within departments.

4. Conclusion

In this paper, I combine new analytical strategies with incident-level data on all recorded uses of force by municipal and state police in New Jersey between 2012 and 2016 to estimate racial disparities in the severity of force used on a subject, conditional on force. This rich dataset allows me to adjust for factors such as the type of incident and the subject's actions and address selection into the sample to better examine the role of race in police use of force. Overall, I find large disparities against Black subjects in police use of force that are larger at higher levels of force. The departments themselves, moreso than their officers or spatial variation in their incidents and subjects, appear to be an important determinant of police use of force. I then extend empirical Bayes methods and document substantial heterogeneity in these racial differences across and within departments, finding some departments without disparities against Black subjects, but also a long tail with especially large ones. A department with a disparity at one level of force is more likely to have a disparity at the others, but many do not, highlighting the complexities of investigating racial inequities in police use of force. I also explore the factors that differentiate departments without disparities against Black subjects from those with the largest ones. The strongest evidence is in favor of municipal characteristics, as opposed to departmental observables, but additional research is required to differentiate correlation and causality.

Much work remains to be done on race and police use of force. The presence of departments without estimated racial disparities is a sign that improvements are possible. But the variation in disparities both across and within departments may make identifying departments for treatment and implementing effective reform more difficult and less efficient. Improved data availability and transparency offer significant promise, and initiatives such as New Jersey's new statewide data portal and the FBI's National Use-of-Force Data Collection will facilitate future studies of how race and policing interact that can better consider how these issues are not uniform, varying across departments, geographies, and force levels. Better understanding of variation in racial disparities in police use of force is a vital step towards greater racial justice and equitable public services.

Acknowledgments

I am grateful to Leah Boustan, Jesse Bruhn, Janet Currie, Milena Djourelova, Felipe Goncalves, Thomas Fujiwara, Bo Honoré, Ilyana Kuziemko, David Lee, Alex Mas, Jonathan Mummolo, Amy Wickett, Jenny Withrow, seminar participants at Princeton University, the U.S. Census Bureau, and the Young Economists Symposium, and members of the Princeton Police Department for helpful comments. This work was primarily conducted while the author was a graduate student at Princeton University and was supported by The Princeton University Industrial Relations Section. Any opinions and conclusions expressed herein are those of the author and do not reflect the views of the U.S. Census Bureau.

Table 1: Summary Statistics for Force Reports

Statistic	N	Mean
Max force: compliance hold	39,321	0.50
Max force: unenhanced	39,321	0.36
Max force: enhanced (non-firearms)	39,321	0.14
Max force: fired weapon	39,321	0.004
Max subject action: resisted	39,321	0.63
Max subject action: physical threat/attack	39,321	0.34
Max subject action: blunt weapon threat/attack	39,321	0.01
Max subject action: knife threat/attack	39,321	0.01
Max subject action: vehicular threat/attack	39,321	0.01
Max subject action: firearm threat	39,321	0.01
Max subject action: fired weapon	39,321	0.001
Officer injured	39,321	0.10
Incident: crime in progress	39,321	0.27
Incident: domestic dispute	39,321	0.13
Incident: other dispute	39,321	0.11
Incident: suspicious person	39,321	0.11
Incident: traffic stop	39,321	0.09
Incident: other	39,321	0.33
Num. Officers Using Force against Subject	39,321	1.65
1 Officer Using Force against Subject	39,321	0.56
2 Officers Using Force against Subject	39,321	0.30
3 Officers Using Force against Subject	39,321	0.10
4 Officers Using Force against Subject	39,321	0.03
5+ Officers Using Force against Subject	39,321	0.01
Subject: White	39,321	0.48
Subject: Black	39,321	0.41
Subject: Hispanic	39,321	0.10
Subject: Asian/Pacific Islander	39,321	0.01
Subject: female	39,321	0.20
Subject: age	39,321	31.09

Notes: Data cover all police departments in New Jersey from 2012 through 2016. Data have been restructured so that each observation is a subject who had force used against them by police and cleaned as described in Appendix A. "Max force" refers to the highest level of force used in an incident; lower levels of force are not reported. "Max subject action" indicates the most severe action a subject took that could justify an office using force. Force used and subject actions are ordered from least severe to most severe. Incidents may have multiple types.

Table 2: Racial Disparities by Police in Force of At Least Specified Severity, Conditional on Force

	Unenhanced		Enhanced		Fired Weapons	
Subject Black	0.019***	0.022***	0.033***	0.034***	0.001	0.001
	(0.005)	(0.006)	(0.005)	(0.005)	(0.001)	(0.001)
Heckman correction		0.071		0.025		0.008*
		(0.045)		(0.031)		(0.004)
Subject female	−0.169***	−0.169***	−0.063***	−0.063***	−0.002***	−0.002***
	(0.007)	(0.007)	(0.005)	(0.005)	(0.000)	(0.000)
Subject age	0.007***	0.007***	0.006***	0.006***	0.000	0.000
	(0.001)	(0.001)	(0.001)	(0.001)	(0.000)	(0.000)
Subject age^2	−0.0001***	−0.0001***	−0.0001***	−0.0001***	−0.000	−0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Resist	−0.273***	−0.273***	−0.195***	−0.195***	−0.006	−0.006
	(0.026)	(0.026)	(0.027)	(0.027)	(0.005)	(0.005)
Physical threat/attack	−0.079***	−0.079***	−0.094***	−0.094***	−0.007	−0.007
	(0.027)	(0.027)	(0.027)	(0.027)	(0.005)	(0.005)
Knife threat/attack	0.031	0.031	0.078**	0.078**	0.063***	0.063***
	(0.038)	(0.038)	(0.038)	(0.038)	(0.017)	(0.017)
Vehicle threat/attack	−0.024	−0.024	−0.051	−0.051	0.084***	0.084***
	(0.035)	(0.036)	(0.033)	(0.033)	(0.026)	(0.026)
Firearm threat	0.000	0.000	0.040	0.040	0.189***	0.189***
	(0.036)	(0.036)	(0.048)	(0.048)	(0.049)	(0.049)
Firearm attack	0.128**	0.128**	0.303***	0.303***	0.576***	0.576***
	(0.051)	(0.051)	(0.074)	(0.074)	(0.079)	(0.079)
Officer harmed	0.122***	0.122***	0.015**	0.015*	0.009***	0.009***
	(0.012)	(0.012)	(0.007)	(0.007)	(0.003)	(0.003)
Fixed effects	Dept.	Dept..	Dept.	Dept..	Dept.	Dept..
Clustering	Dept.	Dept.	Dept.	Dept.	Dept.	Dept.
Outcome mean	0.505	0.505	0.146	0.146	0.004	0.004
R^2	0.182	0.182	0.118	0.118	0.182	0.182
Num. obs.	39183	39183	39183	39183	39183	39183
N Clusters	453	453	453	453	453	453

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$.

Notes: Table reports OLS estimates from Equation 1, which regresses a binary measure of whether police used force of at least the specified severity on indicators for the subject's race, an indicator for the subject being female, a quadratic of the subject's age, time of day fixed effects, indicators for the subject's actions, an indicator for whether an officer was harmed, an indicator for whether the subject was an emotionally disturbed person, indicators for the officer's rank, year fixed effects, and department or officer fixed effects. Some regressions additionally incorporate a Heckman correction-style adjustment adapted from Goncalves and Mello (2021) to examine the possibility of extensive margin racial differences in force usage impacting results. The omitted subject action is threats and attacks with blunt objects.

Table 3: Extended Veil Test of Racial Disparities in Police Use of Force

	No Controls		No Interaction		Controls+Interaction	
K_{evod}	1.572^{*}	1.564^{*}	1.794^{**}	1.703^{*}	1.739^{*}	1.707^{*}
	$(0.953, 2.585)$	$(0.939, 2.618)$	$(1.028, 3.131)$	$(0.960, 3.021)$	$(0.958, 3.158)$	$(0.919, 3.171)$
Intwertwilight Incidents	All	Evening	All	Evening	All	Evening
Controls	N	N	Y	Y	Y	Y
Interact Visibility×Discretion	N	N	N	N	Y	Y
Num. obs.	691	634	691	634	691	634

$^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Notes: Table reports estimates of K_{evod}, the central parameter from the extended veil of darkness test described in Appendix B. Values of K_{evod} greater than 1 are suggestive of extensive margin racial disparities in force usage against Black subjects within this set of intertwilight traffic stops. Columns with all intertwilight incidents include all traffic stops from both morning and evening intertwilight periods, and other columns include on the latter. Results in the first two columns include no additional controls. Results in the third and fourth columns add a control for the proportion of the population ages 18-65 that is Black in the municipality. The final columns further add an interaction between the incident taking place while it is light out and whether the use of force is considered discretionary in my framework rather than assuming their effects are additive. Asymmetric 95% confidence intervals based on exponentiating the log odds confidence interval are in parentheses.

Table 4: Effect of Female Subject on Police Use of Force of At Least Specified Severity, Conditional on Force

	Unenhanced	Enhanced
Subject Black	0.017^{***}	0.033^{***}
	(0.006)	(0.005)
Subject female	-0.173^{***}	-0.062^{***}
	(0.008)	(0.005)
Subject Black × female	0.010	−0.001
	(0.010)	(0.008)
Fixed effects	Dept.	Dept.
Clustering	Dept.	Dept.
Outcome mean	0.505	0.146
R^2	0.182	0.118
Num. obs.	39183	39183

$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Notes: Table reports OLS estimates from Equation 1, which regresses a binary measure of whether police used force of at least the specified severity on indicators for the subject's race interacted with an indicator for the subject being female, a quadratic of the subject's age, time of day fixed effects, indicators for the subject's actions, an indicator for whether an officer was harmed, an indicator for whether the subject was an emotionally disturbed person, indicators for the officer's rank, year fixed effects, and department fixed effects.

Table 5: Officer Harmed Regression

	Officer Harmed
Max Force Unenhanced	0.050^{***}
	(0.006)
Max Force Enhanced	0.040^{***}
	(0.008)
Max Force Fired Weapon	0.156^{***}
	(0.050)
Subject Black	0.018^{***}
	(0.004)
Subject Black × Max Force Unenhanced	0.013^{*}
	(0.008)
Subject Black × Max Force Enhanced	-0.016^{*}
	(0.009)
Subject Black × Max Force Fired Weapon	0.221^{*}
	(0.116)
R^2	0.081
Num. obs.	39183

$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Notes: Table reports OLS estimates based on a modified version of Equation 1 where the outcome is an indicator for whether an officer reported being harmed in that incident. The right-hand side additionally includes indicators for the maximum level of force used and interactions between those force levels and the indicator for the subject being Black.

Table 6: Effect of Being in Department with Above-Median Empirical Bayes Racial Disparity, Officers Switching Departments Only

	Unenhanced	Enhanced
Subject Black	0.024	0.004
	(0.040)	(0.021)
Department above median	-0.022	-0.016
	(0.061)	(0.030)
Subject Black × department above median	0.123^{**}	0.035
	(0.061)	(0.038)
Fixed effects	Off.	Off.
Clustering	Dept.	Dept.
R^2	0.421	0.299
Num. obs.	1095	1095

$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Notes: Table reports OLS estimates from Equation 1, which regresses a binary measure of whether police used force of at least the specified severity on indicators for the subject's race, an indicator for the subject being female, a quadratic of the subject's age, time of day fixed effects, indicators for the subject's actions, an indicator for whether an officer was harmed, an indicator for whether the subject was an emotionally disturbed person, indicators for the officer's rank, year fixed effects, department fixed effects, and an indicator for the department in which the officer served having above-median racial disparities from the empirical Bayes analysis. This regression only includes incidents from officers identified as switching departments.

Table 7: Summary Statistics for Winsorized Empirical Bayes Department x Black Subject Interactions

	Unenhanced	Enhanced	Fired Weapons	Unenhanced Half CI	Enhanced Half CI	Fired Weapons Half CI
SD	0.067	0.051	0.007	2.196	1.463	0.300
Min	-0.220	-0.208	-0.037	0.024	0.009	0.000
P05	-0.045	-0.095	-0.009	0.054	0.023	0.001
P25	0.008	0.001	-0.001	0.164	0.059	0.002
Median	0.019	0.033	0.001	0.425	0.133	0.003
P75	0.031	0.037	0.001	1.094	0.542	0.007
P95	0.104	0.065	0.009	4.822	3.277	0.049
Max	0.370	0.097	0.026	13.472	10.338	2.702
Mean	0.024	0.014	-0.000	1.174	0.668	0.046
% ≤ 0	0.183	0.243	0.369			

Notes: Table reports empirical Bayes estimates and 95% confidence interval half-lengths of departmental racial disparities from Equation 5, which regresses a binary measure of whether police used force of at least the specified severity on indicators for the subject's race interacted with a department indicator, an indicator for the subject being female, a quadratic of the subject's age, time of day fixed effects, indicators for the subject's actions, an indicator for whether an officer was harmed, an indicator for whether the subject was an emotionally disturbed person, indicators for the officer's rank, year fixed effects, and department fixed effects. Winsorization is done at the 1st and 99th percentiles. Parametric confidence intervals use the formulation proposed in Morris (1983).

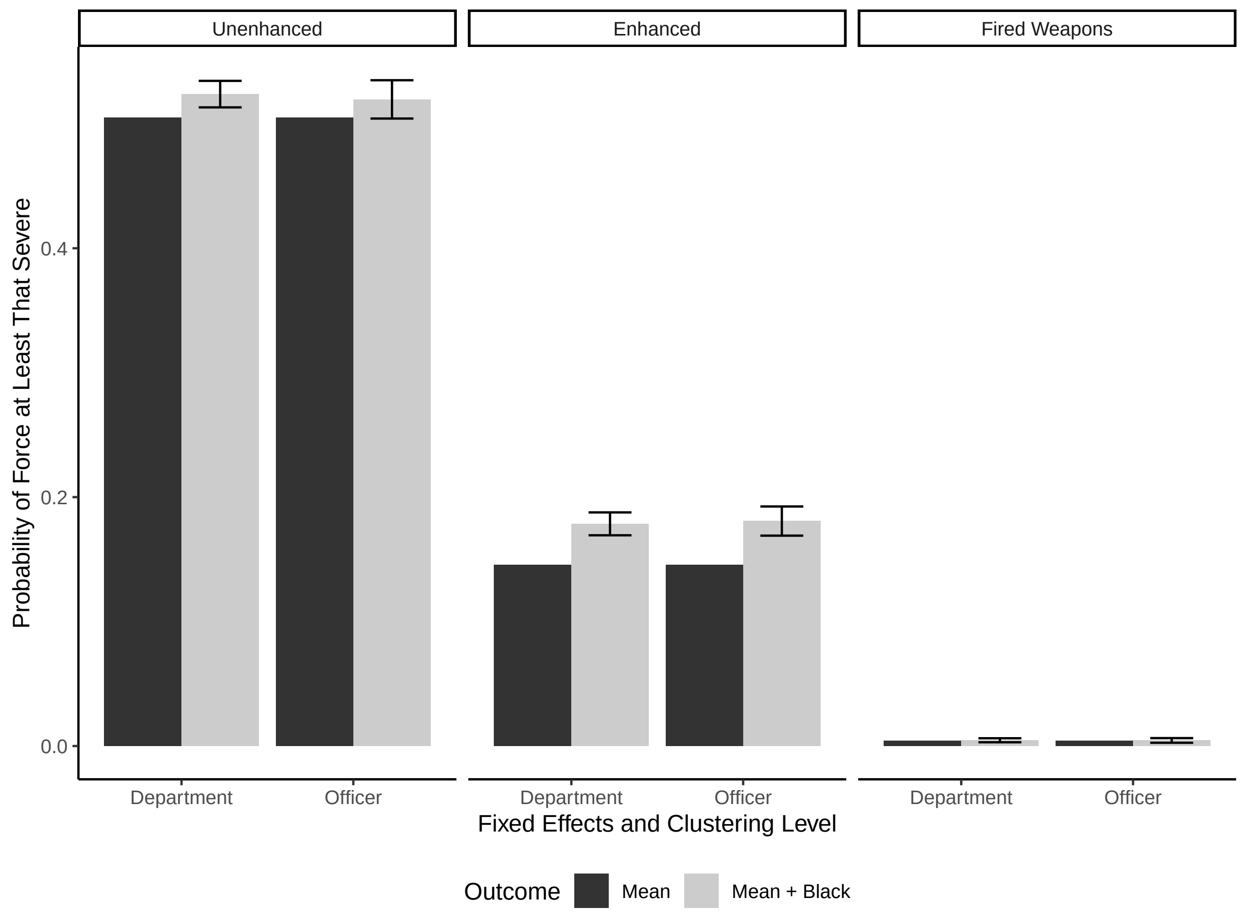

Figure 1: Overall Racial Disparities for Subject Being Black on Probability of Force of at Least Specified Severity
Notes: Figure presents results from a series of OLS models fit via Equation 1. Each heading represents a different outcome: whether, conditional on any force being used, force of at least the specified severity was used. Bars labeled "Department" include department fixed effects, and bars labeled "Officer" instead include officer fixed effects. Confidence intervals are based on the Black coefficient.

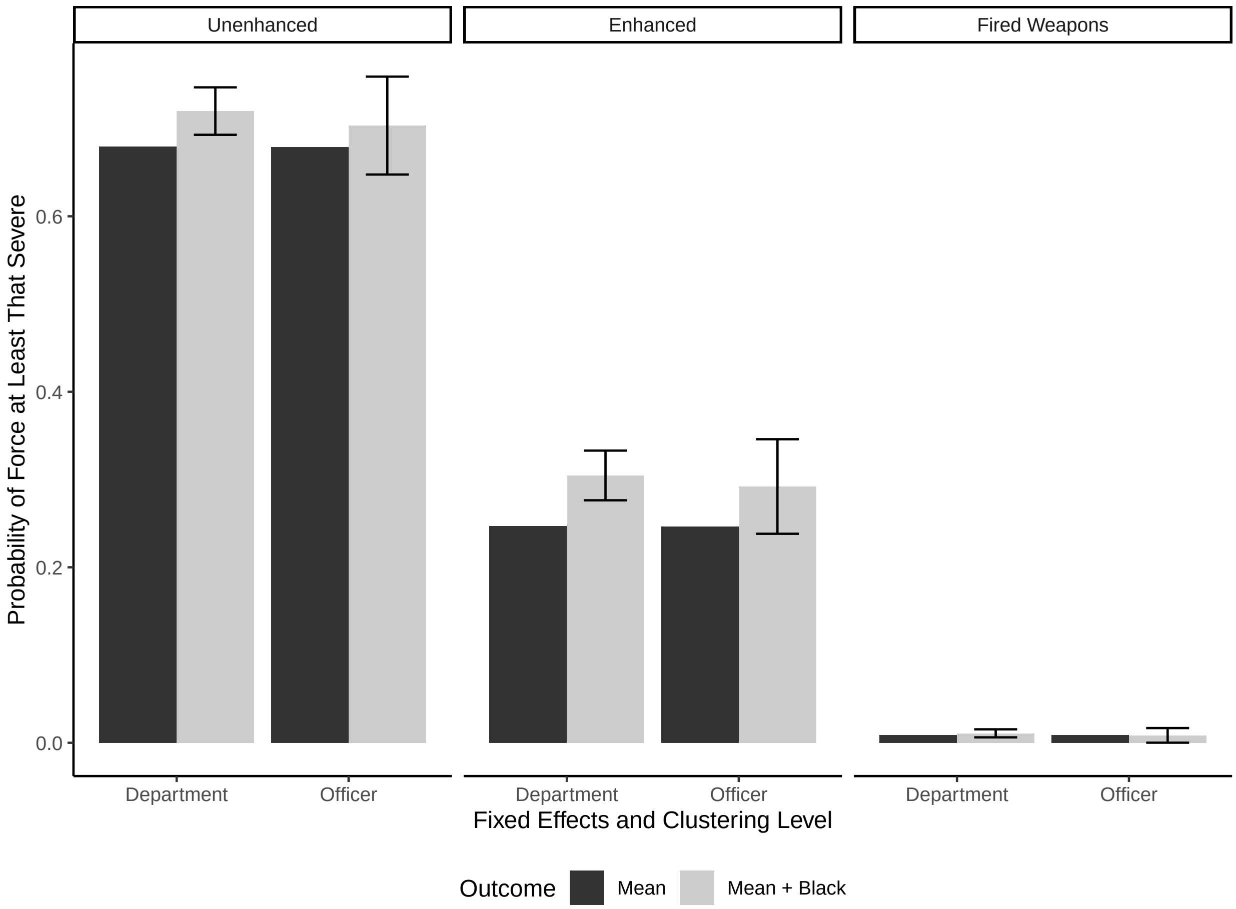

Figure 2: Overall Racial Disparities of Subject Being Black on Probability of Force of at Least Specified Severity (Subset Where Race is Unlikely to Affect Decision to Engage Subject and Subject at Least Physically Threatened/Attacked Officer or Another)

Notes: Figure presents results from a series of OLS models estimated on the subset of the data where a subject's race was less likely to have influenced the officer's decision to engage with the subject: crimes in progress, disputes, and traffic stops at night, and where the subject at least physically threatened or attacked an officer or another. The latter restriction is equivalent to dropping observations where the most severe actions by the subject was resisting. Regressions are fit via Equation 1. Each heading represents a different outcome: whether, conditional on any force being used, force of at least the specified severity was used. Bars labeled "Department" include department fixed effects, and bars labeled "Officer" instead include officer fixed effects. Confidence intervals are based on the Black coefficient.

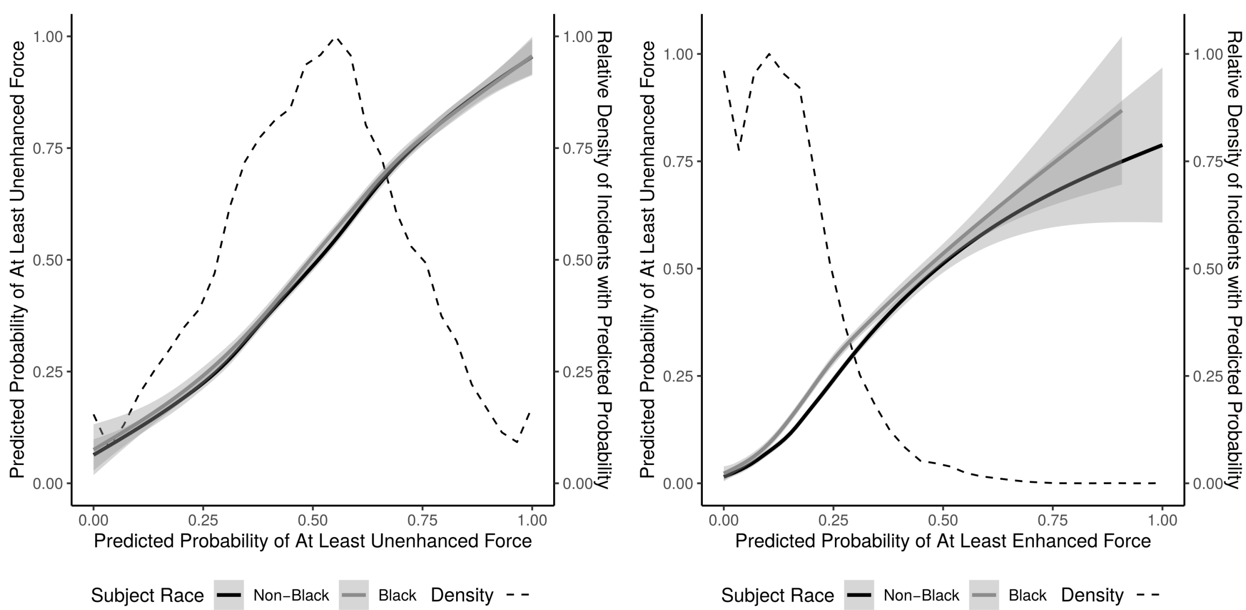

Figure 3: Predicted Probabilities of Force Outcomes: Densities and Relationships with Actual Outcomes, by Race
Notes: Dashed lines indicate the densities of the predicted probabilities of each outcome fit via Equation 1 excluding the racial indicator. Solid lines plot these predicted probabilities against the actual proportions of incidents with that outcome, by race, with 95% confidence intervals shaded. Outcomes are smoothed via cubic splines. Predicted probabilities are censored at 0 and 1 to prevent overfitting on small numbers of observations with predicted probabilities outside of that range (1.8% of incidents for at least unenhanced and 8.4% for enhanced).

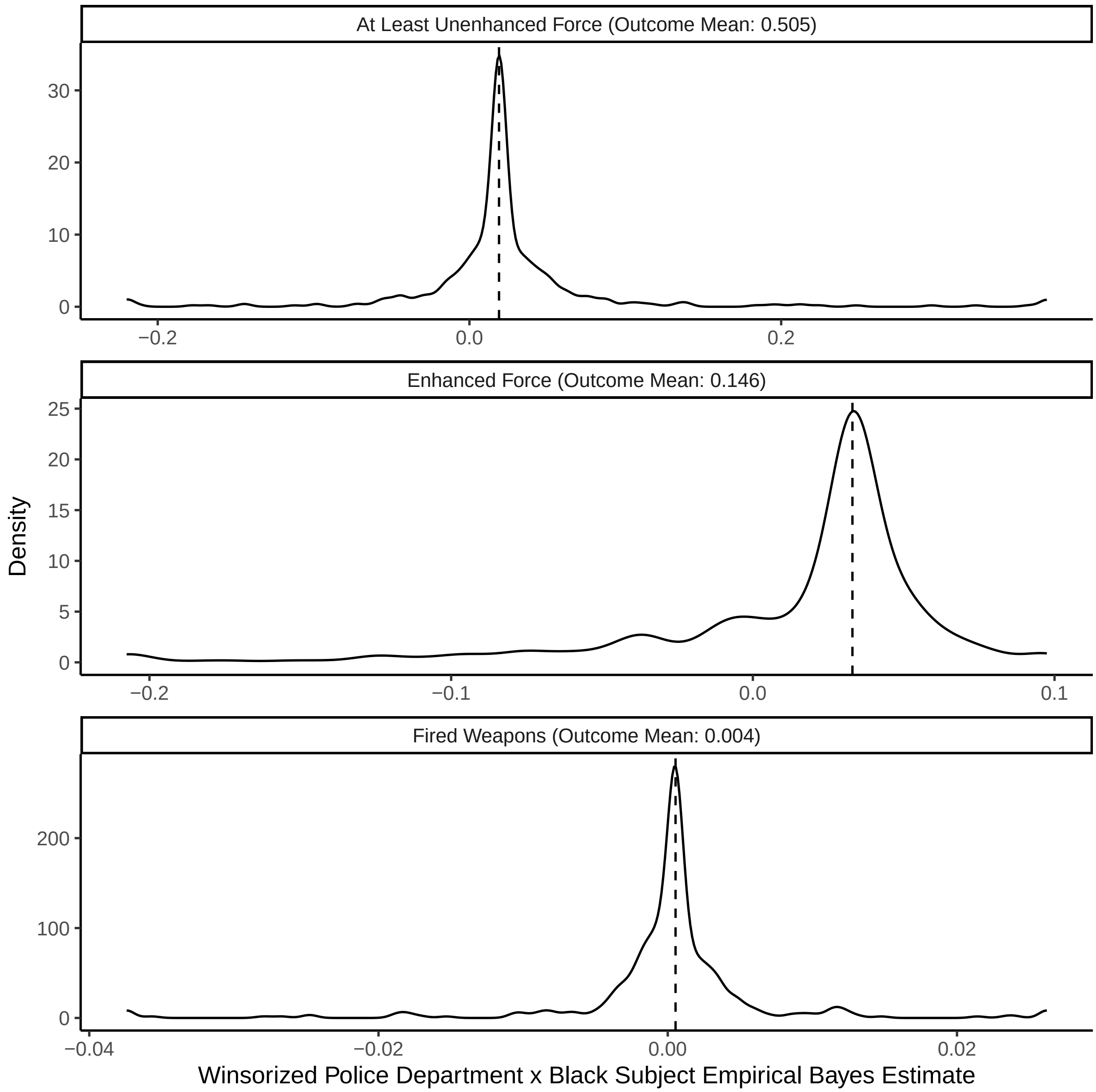

Figure 4: Distribution of Winsorized Empirical Bayes Estimates of Department-Black Interactions

Notes: Figure presents kernel density estimates of department-specific racial disparities β between White/Asian/Pacific Islander subjects and Black subjects as estimated from the empirical Bayes estimator in Equation 5 and winsorized at the 1% and 99% levels with Gaussian kernels and the Silverman (1986) rule-of-thumb bandwidth. Each subgraph shows results from regressions with the specified outcome outcome: whether, conditional on any force being used, force of at least the specified severity was used. The dashed line indicates the mean of the prior.

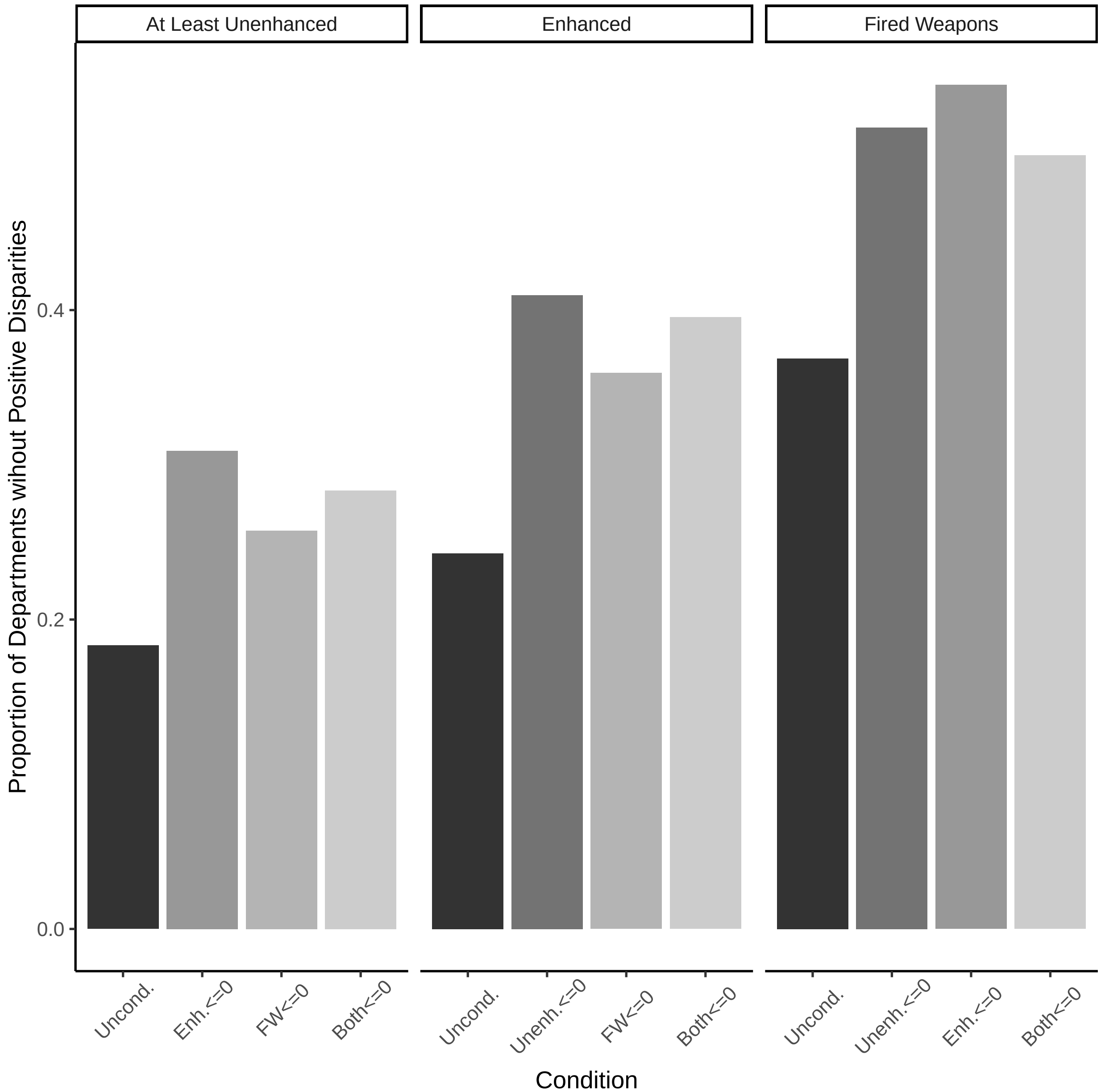

Figure 5: Within-Department Empirical Bayes Estimates and Department-Black Interaction Coefficients Across Force Types

Notes: Figure presents conditional probabilities that a given department does not have a disparity in its use of force against Black subjects. For each of the three outcomes (at least unenhanced, enhanced, and fired weapons), there are four bars: one for the unconditional probability that a department does not have a disparity against Black subjects, two for the probabilities conditional on each of the other outcomes not having a disparity, and one for the probability conditional on both other outcomes not having disparities.

References

Albert, Adelin and John A Anderson (1984). "On the Existence of Maximum Likelihood Estimates in Logistic Regression Models". *Biometrika* 71.1, pp. 1–10.

Ang, Desmond, Panka Bencsik, Jesse Bruhn, and Ellora Derenoncourt (2021). *Police Violence Reduces Civilian Cooperation and Engagement with Law Enforcement*. Working Paper 2021-005. Orlando Bravo Center for Economic Research. URL: `https://economics.brown.edu/sites/g/files/dprerj726/files/papers/Bravo%20Working%20Paper%202021-005.pdf`.

Antonovics, Kate and Brian G Knight (2009). "A New Look at Racial Profiling: Evidence from the Boston Police Department". *The Review of Economics and Statistics* 91.1, pp. 163–177.

Anwar, Shamena and Hanming Fang (2006). "An Alternative Test of Racial Prejudice in Motor Vehicle Searches: Theory and Evidence". *American Economic Review* 96.1, pp. 127–151.

Arnold, David, Will Dobbie, and Crystal S. Yang (2018). "Racial Bias in Bail Decisions". *The Quarterly Journal of Economics* 133.4, pp. 1885–1932.

Ba, Bocar, Dean Knox, Jonathan Mummolo, and Roman Rivera (2021). "The Role of Officer Race and Gender in Police-Civilian Interactions in Chicago". *Science* 371.6530, pp. 696–702.

Becker, Gary S (1957). *The Economics of Discrimination*. University of Chicago Press.

Bielen, Samantha, Wim Marneffe, and Naci H. Mocan (2019). *Racial Bias and In-group Bias In Judicial Decisions: Evidence from Virtual Reality Courtrooms*. Working Paper 25355. National Bureau of Economic Research. URL: `http://www.nber.org/papers/w25355`.

Bohren, J. Aislinn, Kareem Haggag, Alex Imas, and Devin G. Pope (2020). *Inaccurate Statistical Discrimination: An Identification Problem*. Tech. rep. National Bureau of Economic Research.

Bureau of Justice Statistics (2016). *Law Enforcement Management and Administrative Statistics (LEMAS)*. DOI: `10.3886/ICPSR37323.v1`.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff (2014). "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates". *American Economic Review* 104.9, pp. 2593–2632.

Coviello, Decio and Nicola Persico (2015). "An Economic Analysis of Black-White Disparities in the New York Police Department's Stop-and-Frisk Program". *The Journal of Legal Studies* 44.2, pp. 315–360.

Cunningham, Jamein, Donna Feir, and Rob Gillezeau (2021). *Collective Bargaining Rights, Policing, and Civilian Deaths*. Tech. rep. URL: `http://ftp.iza.org/dp14208.pdf`.

Davis, Elizabeth, Anthony Whyde, and Lynn Langton (2018). "Contacts between police and the public, 2015". *US Department of Justice Office of Justice Programs Bureau of Justice Statistics Special Report*, pp. 1–33.

Devi, Tanaya and Roland G. Fryer (2020). *Policing the Police: The Impact of "Pattern-or-Practice" Investigations on Crime*. Working Paper 27324. National Bureau of Economic Research. URL: `http://www.nber.org/papers/w27324`.

Durlauf, Steven N. and James J. Heckman (2020). "An Empirical Analysis of Racial Differences in Police Use of Force: A Comment". *Journal of Political Economy* 128.10, pp. 3998–4002.

Fryer, Roland G. (2019). "An Empirical Analysis of Racial Differences in Police Use of Force". *Journal of Political Economy* 127.3, pp. 1210–1261.

— (2020). "An Empirical Analysis of Racial Differences in Police Use of Force: A Response". *Journal of Political Economy* 128.10, pp. 4003–4008.

Gelman, Andrew, Jeffrey Fagan, and Alex 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, pp. 813–823.

Goncalves, Felipe and Steven Mello (2021). “A Few Bad Apples? Racial Bias in Policing”. *American Economic Review* 111.5, pp. 1406–41.

Goodman-Bacon, Andrew (2021). “Difference-in-differences with variation in treatment timing”. *Journal of Econometrics* 225.2, pp. 254–277.

Grogger, Jeffrey and Greg Ridgeway (2006). “Testing for Racial Profiling in Traffic Stops from Behind a Veil of Darkness”. *Journal of the American Statistical Association* 101.475, pp. 878–887.

Gu, Jiaying and Chris Walters (2022). *Empirical Bayes Methods: Theory and Application*. URL: http://conference.nber.org/confer/2022/SI2022/ML/Slides2.pdf.

Hamermesh, Daniel S (1996). “Workdays, workhours and work schedules: evidence for the United States and Germany”.

Heaton, Paul (2010). “Understanding the Effects of Antiprofiling Policies”. *The Journal of Law and Economics* 53.1, pp. 29–64.

Heckman, James (1990). “Varieties of Selection Bias”. *The American Economic Review* 80.2, pp. 313–318.

Hoekstra, Mark and CarlyWill Sloan (2020). *Does Race Matter for Police Use of Force? Evidence from 911 Calls*. Working Paper 26774. National Bureau of Economic Research. URL: http://www.nber.org/papers/w26774.

Horrace, William C. and Shawn M. Rohlin (2016). “How Dark Is Dark? Bright Lights, Big City, Racial Profiling”. *The Review of Economics and Statistics* 98.2, pp. 226–232.

Kane, Thomas J. and Douglas O. Staiger (2008). “Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation”. National Bureau of Economic Research working paper 14607.

Kaplan, Jacob (2021). *Jacob Kaplan’s Concatenated Files: Uniform Crime Reporting (UCR) Program Data: Arrests by Age, Sex, and Race, 1974-2020*. Ann Arbor, MI: Inter-university Consortium for Political and Social Research. URL: https://doi.org/10.3886/E102263V14.

Kline, Patrick, Evan K Rose, and Christopher R Walters (2022). “Systemic discrimination among large US employers”. *The Quarterly Journal of Economics* 137.4, pp. 1963–2036.

Knowles, John, Nicola Persico, and Petra Todd (2001). “Racial Bias in Motor Vehicle Searches: Theory and Evidence”. *Journal of Political Economy* 109.1, pp. 203–229.

Knox, Dean, Will Lowe, and Jonathan Mummolo (2020). “Administrative Records Mask Racially Biased Policing”. *American Political Science Review* 114.3, pp. 619–637.

Luh, Elizabeth (2020). *Not So Black and White: Uncovering Racial Bias from Systematically Misreported Trooper Reports*. Tech. rep. URL: https://elizluh.github.io/files/highwaybias.pdf.

MacKinnon, James G, Morten Ørregaard Nielsen, and Matthew D Webb (2022). *Leverage, Influence, and the Jackknife in Clustered Regression Models: Reliable Inference Using Summclust*. Tech. rep. URL: https://arxiv.org/abs/2205.03288.

McCarthy, Craig (2019). *Help us continue to improve N.J’s first statewide database of police use of force*. URL: https://www.nj.com/news/2018/12/help-us-continue-to-improve-njs-first-statewide-database-of-police-use-of-force.html (visited on 06/01/2019).

McCarthy, Craig and Blake Nelson (2019). *Every way a police officer is legally allowed to harm another person, ranked*. URL: https://www.nj.com/news/2018/11/every_way_a_police_officer_is_legally_allowed_to_harm_another_person_from_a_to_z.html (visited on 06/01/2019).

McCarthy, Craig and Stephen Stirling (2019). *How we built the most comprehensive statewide database of police force in the United States*. URL: https://www.nj.com/news/2018/11/how_we_built_the_most_comprehensive_statewide_database_of_police_force_in_the_us.html (visited on 06/01/2019).

Morris, Carl N. (1983). "Parametric Empirical Bayes Inference: Theory and Applications". *Journal of the American Statistical Association* 78.381, pp. 47–55.

Nelson, Blake (2019). *Frequently asked questions about The Force Report*. URL: https://www.nj.com/news/2018/11/frequently_asked_questions_about_the_force_report.html (visited on 06/01/2019).

New Jersey Attorney General (2000). *Use of Force*. URL: https://www.nj.gov/oag/dcj/agguide/useofforce2001.pdf.

Nicholson-Crotty, Sean, Jill Nicholson-Crotty, and Sergio Fernandez (2017). "Will More Black Cops Matter? Officer Race and Police-Involved Homicides of Black Citizens". *Public Administration Review* 77.2, pp. 206–216.

Nix, Justin, Bradley A Campbell, Edward H Byers, and Geoffrey P Alpert (2017). "A Bird's Eye View of Civilians Killed by Police in 2015: Further Evidence of Implicit Bias". *Criminology & Public Policy* 16.1, pp. 309–340.

NJ Advance Media (2019). *The Force Report*. URL: http://force.nj.com (visited on 06/01/2019).

Persico, Nicola and Petra Todd (2006). "Generalising the Hit Rates Test for Racial Bias in Law Enforcement, with an Application to Vehicle Searches in Wichita". *The Economic Journal* 116.515, F351–F367.

Raphael, Steven and Sandra V. Rozo (2019). "Racial Disparities in the Acquisition of Juvenile Arrest Records". *Journal of Labor Economics* 37.S1, S125–S159.

Rehavi, M. Marit and Sonja B. Starr (2014). "Racial Disparity in Federal Criminal Sentences". *Journal of Political Economy* 122.6, pp. 1320–1354.

Ridgeway, Greg and John MacDonald (2010). "Methods for Assessing Racially Biased Policing". In: *Race, Ethnicity, and Policing: New and Essential Readings*. Ed. by Stephen K. Rice and Michael D. White. New York University Press, pp. 180–204.

Ross, Cody T. (2015). "A Multi-Level Bayesian Analysis of Racial Bias in Police Shootings at the County-Level in the United States, 2011–2014". *PLoS One* 10.11, e0141854.

Ruggles, Steven, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas, and Matthew Sobek (2020). *IPUMS USA: Version 10.0 [dataset]*. Minneapolis, MN. URL: https://doi.org/10.18128/D010.V10.0.

Silverman, Bernard W. (1986). *Density Estimation for Statistics and Data Analysis*. Vol. 26. CRC Press.

Tuttle, Cody (2021). *Racial Disparities in Federal Sentencing: Evidence from Drug Mandatory Minimums*. Tech. rep. URL: https://codytuttle.github.io/tuttle_mandatory_minimums.pdf.

Weisburst, Emily K. (2019). "Police Use of Force as an Extension of Arrests: Examining Disparities across Civilian and Officer Race". *AEA Papers and Proceedings* 109, pp. 152–56.

West, Jeremy (2018). *Racial Bias in Police Investigations*. Tech. rep. URL: https://people.ucsc.edu/~jwest1/articles/West_RacialBiasPolice.pdf.

Appendix (For Online Publication Only)

Appendix A. Data Cleaning

Despite the substantial efforts by the teams at ProPublica and NJ Advance Media, the use of force dataset requires additional processing before it can be be used for analysis. Here I outline the changes I make to the data.

The structure of each department's force reports and how many officers or subjects can be put on a single report vary. For consistency, I structure the data so that every observation is one subject in one incident who has force used against them by one officer. When there are multiple force reports for a single subject in an incident, I identify duplicates by computing similarity scores based on the Levenshtein distance between subject names, preventing situations such as a missing middle initial in one report or typos from creating repeat incidents. For 167 observations where multiple officers are recorded on the same force report without differentiating between their actions, I use the first officer listed. Because my empirical strategies are based on the most severe level of force used against a subject, for incidents where multiple officers use force against a single subject, I keep the officer who uses the greatest level of force, choosing randomly in the event of ties.

Besides restructuring, the most significant modifications I make to the data are for the types of force used. These are text strings in the raw data and contain many irregularities. Some have typos, some do not directly correspond with force categories (for example, "Grabbed Gardening Tool Out Of Her Hand"), and some forms, notably from the New Jersey State Police, have their own names for force levels, such as "physical" and "mechanical." Many of these force classifications require subjective judgments, which I base on officer narratives whenever possible. For instances where an officer uses a firearm as a blunt weapon ("pistol whipping"), I mark this as a use of a baton, the most similar type of force. Similarly, when blunt objects such as flashlights are used to strike a subject, I record them as batons. When batons are used for leverage in compliance holds, I count these incidents as compliance holds. I do not count restraints such as handcuffs or "the wrap" as force, but maneuvers to facilitate them may be, such as forcing one's arms behind their back. Unless there is an indication that a subject was punched, slapped, or otherwise struck, I classify hands on the subject as compliance holds rather than the use of hands/fists. If an officer pushes a subject to get them to move, such as pushing them into a police vehicle, I classify it as a compliance hold. When the pushing is done to incapacitate a subject, such as pushing them off of a bicycle, I classify it as a use of hands/fists. I classify forcibly moving actively resisting subjects as compliance holds, but do not count moving passively resisting subjects, such as when a subject sits down and does not move.

Two other variables require manual cleaning: the unique officer identifier and the actions of the subject. The officer identifier variable present in the raw dataset does not track officers as they move across departments. Using available information on race, experience, rank, and geographic location, I create an officer identifier variable that follows officers across departments. The subject actions variable is structured in the same manner as the force used variable and suffers from the same inconsistency problems. Again I manually map this variable to the maximum action a subject took that may have justified force. The lowest category is resisting, which includes physically resisting an officer's control, fleeing officer apprehension, and actions that do not fit into another category. The next categories, physical, blunt, knife, and vehicle, cover both attacks and threats of attack. I split incidents involving firearms into threats with firearms and actually firing them. I only count subject actions directed at humans, ignoring behaviors such as subjects attempting to kick out the windows of the police vehicle in which they are being held. I consider flailing and spitting or using other bodily fluids as projectiles to be a physical attack. When a subject attempts to disarm an officer, I mark it as a physical attack.

For the incident type variable, I use the categories present on the New Jersey Attorney General's model form (Figure A.1): crime in progress, domestic, other dispute, suspicious person, and other, also using "other" for incidents that do not fit into another category. Some incidents have multiple types, which I allow, except for the "other" category.

I remove a small number of irregular or non-representative observations from the data. I drop 963 observations (more than 600 of them would have been dropped anyway due to other missing variables) where no reason for the force was given, as it is more likely that the officer-stated reason is missing than nonexistent. I remove a handful of incidents involving nonlethal firearms ("riot guns"), as these are not standard police carry. I do not include the constructive authority category of force, incidents where a firearm or stun gun is drawn but not fired to induce compliant behavior, as no physical force is used in these situations. I remove 44 individuals whose indicated races do not fall within the categories of White, Black, Hispanic, or Asian/Pacific Islander, such as people marked as "mixed."

Appendix B. Extended Veil of Darkness Test

Appendix B.1. Model

The veil of darkness test developed by Grogger and Ridgeway (2006) elegantly outlines conditions to test for racial bias in motor vehicle stops. I further adapt this test to examine racial differences in police use of force. Let F denote that a subject has force used against them by police and B ($\overline{B}$) denote that a subject is (not) Black and at risk of having force used against them by police. The ideal, unfeasible veil of darkness test from Grogger and Ridgeway (2006) for traffic stops centers around incidents where the officers can(not) observe the subject's race prior to their decision to stop them, V ($\overline{V}$). To examine use of force within this framework, suppose that we could also "veil" the subject's race from the extensive margin decision of whether to use force at all after the initial stop. Although the subject's race is virtually always going to be observable to the officer when making this decision, consider incidents where that decision is (not) discretionary, D ($\overline{D}$). For example, if the subject attacks an officer, we can expect that police will respond with force no matter what the race of the subject is, effectively removing the effects of race on the extensive margin within these incidents and "veiling" them. This leads to the ideal extended veil test, which follows directly from replacing V in the original veil test with V and D.

$$\frac{P(F|V,D,B)}{P(F|V,D,\overline{B})} = K_{ideal} \times \frac{P(F|\overline{V},\overline{D},B)}{P(F|\overline{V},\overline{D},\overline{B})} \tag{B.1}$$

The left-hand side is the relative risk of a Black subject of having force used against them when their race is observable before stopping them and the force usage is discretionary. The right-hand side has the analogous risk when race is not observable before the stop and the force is non-discretionary, as well as a multiplicative factor K_{ideal}. In the absence of racial differences in stoppage and force decisions, $K_{ideal} = 1$, or the extent to which $K_{ideal} \neq 1$ gives a measure of those gaps.

Like the ideal veil test from Grogger and Ridgeway (2006), the ideal extended veil test is infeasible, as it depends on unobservable quantities. To make it tractable, start by applying Bayes' rule to Equation B.1.

$$K_{ideal} = \frac{P(B|F,V,D)P(\overline{B}|F,\overline{V},\overline{D})}{P(\overline{B}|F,V,D)P(B|F,\overline{V},\overline{D})} \times \frac{P(B|\overline{V},\overline{D})P(\overline{B}|V,D)}{P(\overline{B}|\overline{V},\overline{D})P(B|V,D)} \tag{B.2}$$

The first term is based on the ratio of Black force subjects to others when an incident is and is not double-veiled. Under an assumption of constant relative risk, the ratio of Black to other suspects is independent of daylight and darkness, and the second line equals one and can be ignored. Because V is unobserved, Grogger and Ridgeway (2006) use a darkness indicator $visible$ to indicate whether the subject's race is less likely to be observable before stopping them. For the extensive margin force decision, instead of D, consider an indicator $discretionary$ based on whether the subject took actions at least as severe as physically threatening the officer or another individual, which makes extensive margin force usage less discretionary. Substituting these into Equation B.2 then yields

$$
\begin{aligned}
K' = & \frac{P(B|F, visible=1, discretionary=1)P(\overline{B}|F, visible=0, discretionary=0)}{P(\overline{B}|F, visible=1, discretionary=1)P(B|F, visible=0, discretionary=0)} \times \\
& \frac{P(B|visible=0, discretionary=0)P(\overline{B}|visible=1, discretionary=1)}{P(\overline{B}|visible=0, discretionary=0)P(B|visible=1, discretionary=1)}.
\end{aligned} \tag{B.3}
$$

Under a constant relative risk assumption, the second line equals one, and we get the extended veil of darkness test.

$$
K_{evod} = \frac{P(B|F, visible=1, discretionary=1)P(\overline{B}|F, visible=0, discretionary=0)}{P(\overline{B}|F, visible=1, discretionary=1)P(B|F, visible=0, discretionary=0)} \tag{B.4}
$$

Grogger and Ridgeway (2006) outline three assumptions for the validity of the veil of darkness test, which are straightforward to move into this extended veil test.

1. $K_{ideal} > 1$.
2. $P(V \cap D|visible=1, discretionary=1) > P(V \cap D|visible=0, discretionary=0)$
3. $\frac{P(\overline{B}|visible=1, discretionary=1)P(B|visible=0, discretionary=0)}{P(B|visible=1, discretionary=1)P(\overline{B}|visible=0, discretionary=0)} = 1$

Assumption 1 is that there is a racial disparity against Black subjects in force usage. Assumption 2 is a generalization of the assumption that darkness has a race-blinding effect, $P(V|visible = 1) > P(V|visible = 0)$. Here, we also assume that the indicator for the subject at least physically threatening or attacking someone increases the probability that the use of force was not discretionary. Note that the assumption is not that *visible* and *discretionary* perfectly correlate with V and D. Assumption 3 is the constant relative risk assumption. This is a strong assumption, but it can be relaxed by focusing only on intertwilight periods, clock times that are both dark or light at different times depending on daylight savings time and the tilt of the Earth, as well as by controlling for clock time directly when implementing the test. If these three assumptions hold, then $1 < K_{evod} \leq K_{ideal}$. The test is qualitative. It can detect racial disparities, but interpreting the magnitude of K_{evod} as if it were K_{ideal} would require a much stronger version of Assumption 2: that $P(V \cap D|visible = 1, discretionary = 1) = 1$ and $P(V \cap D|visible = 0, discretionary = 0) = 0$.

The extended veil test is simple to execute with logistic regression.

$$
\frac{P(B|F, visible, discretionary, t)}{1 - P(B|F, visible, discretionary, t)} = \exp(\beta_1 visible + \beta_2 discretionary + \gamma' s(t)) \tag{B.5}
$$

Note the addition of a function of continuous clock time t; I use separate cubic splines for the morning and evening intertwilight periods. With this formulation, the test statistic of interest is $\log(K_{evod}) = -\beta_1 - \beta_2$. The specification is flexible. Following Grogger and Ridgeway (2006) and Horrace and Rohlin (2016), I also add as a control in some models the responding department's percentage of the population ages 18-65 that is Black to help ensure the robustness of the constant relative risk assumption. I further test adding an interaction between *visible* and *discretionary* rather than assuming that their effects are additive.

Appendix B.2. Sensitivity and Robustness

The extended veil test, like the veil test on which it is based, depends on the constancy of the Black/not-Black risk ratios. In practice, it may also depend on reporting rates. The extended veil test is robust to nonreporting if the proportion of missingness is constant within subject race across *visible* and *discretionary*. While these factors are not observable, Grogger and Ridgeway (2006) outline several checks to examine the sensitivity of the veil test's findings by asking how large the risk ratio differences or non-reporting problems would have to be to change findings.

To understand the magnitude of the biases necessary to change the extended veil test's results, suppose there is no nonreporting bias and consider the third column of Table 3. The estimate of interest is $\log(K_{evod}) = -\beta_1 - \beta_2 = 0.585$, and the lower bound of a 90% confidence interval for that linear combination is 0.119. To reverse the positive point estimate, the constant relative risk assumption would need to be violated such that

$$\frac{\frac{P(\overline{B}|visible=1,discretionary=1,t)}{P(B|visible=1,discretionary=1,t)}}{\frac{P(\overline{B}|visible=0,discretionary=0,t)}{P(B|visible=0,discretionary=0,t)}} = \exp(-0.585) = 0.557. \tag{B.6}$$

Suppose that at a given intertwilight time t on a day when it is light, Black subjects are 40.6% of the at-risk population in discretionary incidents (i.e., $P(B|visible = 1, discretionary = 1, t) = 0.406$), matching the overall sample mean. In that case, reversing the sign of the point estimate would require that Black subjects make up 27.5% of the at-risk population at that time in non-discretionary incidents when it is dark at t, a 32% decrease (it would require an 7% decrease for the lower bound of the 90% confidence interval to be below 1).

Next, consider nonreporting, assuming that there is no violation of the risk ratio assumption. Let R denote a force incident was reported, as is required by law. To change the sign of the point estimate, we would need

$$\frac{\frac{P(R|F,B,visible=0,discretionary=0,t)}{P(R|F,B,visible=1,discretionary=1,t)}}{\frac{P(R|F,\overline{B},visible=0,discretionary=0,t)}{P(R|F,\overline{B},visible=1,discretionary=1,t)}} = \exp(-0.585) = 0.557. \tag{B.7}$$

If the reporting rate for non-Black drivers changes with clock time t but not by *visible* and *discretionary*, then the denominator reduces to 1. In this case, force incidents involving Black drivers would have to be 44.3% less likely to be reported when it is dark and the subject at least physically threatens or attacks someone than when neither of these holds, which strains credibility. Even to change the sign on the lower bound of the 90% confidence interval (which would need an 11.2% decrease) seems implausible, as there is no evidence of widespread nonreporting in the force reports, and these non-discretionary incidents are more severe, so the subject actions required to be considered non-discretionary in my framework would likely justify the use of some level of force by the officer.

Appendix C. Heckman Correction for Selection into the Data

Appendix C.1. Model

The force reports do not contain incidents where force might have been used but was not. Depending on correlations with race, this extensive margin of force can affect my estimates of racial disparities along the intensive margin. Here, I outline an approach to adjust for this possibility adapted from Appendix F of Goncalves and Mello (2021) and similar to a Heckman correction with modifications due to the data only having incidents involving force.

Imagine a two-step model of force usage. For narrative simplicity, consider officers as homogeneous within their departments and ignore covariates (my actual implementation of this strategy includes all covariates from the main model). First, an officer from department j must decide whether to use force against a subject of race i at all, the latent variable Z_{ij} that encompasses all elements that affect the extensive margin of force.

$$Z_{ij} = \alpha_j^{NonBlack} + \alpha_j^{Black} \cdot Black_i + \eta_{ij}$$

If $Z_{ij} \geq 0$, then the officer decides to use force against the subject, and chooses the level of force D_{ij}^* to use against the subject.

$$D_{ij}^* = \theta_j^{NonBlack} + \theta_j^{Black} \cdot Black_i + \varepsilon_{ij}$$

D_{ij}^* is a latent variable. In the force reports, we actually observe:

$$D_{ij} = \begin{cases} \mathbb{1}(D_{ij}^* \geq 0) & \text{if } Z_{ij} \geq 0 \\ \text{Missing} & \text{if } Z_{ij} < 0 \end{cases}$$

Assume for narrative consistency that D_{ij}^* and D_{ij} are indicators for whether force of at least a certain severity was used. While D_{ij}^* can be 0 for all levels of force if no force was used in an incident, D_{ij} is always 1 when considering the at least compliance hold measure, because the data only contain incidents where force was used ($Z_{ij} \geq 0$). Consequentially, the racial differences θ that I estimate are based only on observed incidents.

$$\begin{aligned} \hat{\theta}_j^{Black} &= E[D_{ij}^* | Black_i = 1, Z_{ij} > 0] - E[D_{ij}^* | Black_i = 0, Z_{ij} > 0] \\ &= \theta_j^{Black} + E[\varepsilon_{ij} | \eta_{ij} > -\alpha_j^{NonBlack} - \alpha_j^{Black}] - E[\varepsilon_{ij} | \eta_{ij} > -\alpha_j^{NonBlack}] \end{aligned}$$

Under this two-stage model of selection into the sample, there can be problems where the observed subjects are not comparable across races. There are two requirements for this to occur. First, there must of course be a racial difference in the extensive margin of force: $\alpha_j^{Black} \neq 0$. Second, there must also be a correlation in the residuals of the two stages: $corr(\varepsilon_{ij}, \eta_{ij}) \neq 0$. Under those conditions, my estimate of θ_j^{Black} is inconsistent. Unlike in Goncalves and Mello (2021), because I am not strictly estimating a causal treatment effect, these estimates would still be interpretable as the race gap in force experienced conditional on force being used, but this moves away from my target parameter of the difference in force experienced solely from this intensive margin of force with Black and non-Black subjects directly comparable (after adjusting for incident factors, which could also help to correct for these differences).

Suppose now that arrests represent an appropriate "denominator" for use of force, i.e., that they represent the set of incidents in which in officers might use force. We can define the department-race-specific probability that a subject has force used against them as

$$P_{ij} \equiv P(Z_{ij} = 1) = \frac{\#\text{Force}_{ij}}{\#\text{Arrests}_{ij}}.$$

As shown in Goncalves and Mello (2021), it follows that P_{ij} can be used to compute the expectation of the error term η_{ij} for subjects who have force used against them. Assuming $\eta_{ij} \sim N(0,1)$, we get

$$\begin{aligned} P_{ij} &= P(\alpha_j^{NonBlack} + \alpha_j^{Black} \cdot Black_i + \eta_{ij} \geq 0) \\ &= \Phi(\alpha_j^{NonBlack} + \alpha_j^{Black} \cdot Black_i) \\ \implies E(\eta_{ij} | Z_{ij} = 1) &= \frac{\phi(\alpha_j^{NonBlack} + \alpha_j^{Black} \cdot Black_i)}{\Phi(\alpha_j^{NonBlack} + \alpha_j^{Black} \cdot Black_i)} \\ &= \frac{\phi(\Phi^{-1}(P_{ij}))}{P_{ij}}. \end{aligned}$$

One can then compute the expectation of this selection bias in the intensive margin disparities.

$$\begin{aligned} E(\varepsilon_{ij} | \eta_{ij} > -\alpha_j^{NonBlack} - \alpha_j^{Black} \cdot Black_i) &= \rho \cdot E[\eta | \eta_{ij} > -\alpha_j^{NonBlack} - \alpha_j^{Black} \cdot Black_i] \\ &= \rho \cdot E[\eta_{ij} | Z_{ij} = 1] \end{aligned}$$

Adding the expectation of this term (the Mills ratio) to my regressions then adjusts for possible selection bias. An insignificant estimate of the ρ coefficient on the Mills ratio and no change in the estimated disparities or other coefficients, as I find in Table 2, suggest minimal impact of any such bias on my estimates.

Appendix C.2. Arrests as a Benchmark for the At-Risk Population

Arrests are commonly used as a statistical benchmark in assessing racial bias in policing (Ridgeway and MacDonald, 2010). Within the context of police use of force, they may provide a reasonable proxy for the number of incidents where force might have been used, whether or not it actually was. This assumption is used in the Heckman correction estimates and also for the ordered logit models that include FBI arrests data. While this is not falsifiable—if we knew the appropriate number, we would not need benchmarking—we can examine the relationship between force and arrests to see if this is plausible.

It seems reasonable to suppose that all arrests have the potential for escalation to police use of force, as they involve close contact between officers and civilians in an adversarial setting. Ideally, an arrest in an incident would also be necessary for force to be used. If officers often use force without making an arrest, then arrests are likely a poor measure for the population at-risk of having force used against them. The force reports dataset made available to researchers contains a field for charges for subjects (but not a direct arrests field), leading to the natural question of what proportion of force incidents have nonmissing charges data. Charges without an arrest are technically possible, as with warrants or citations, but these court-ordered situations would not appear in the force data. Although it is possible for officers to generate force reports without arresting and charging a subject, these situations would be atypical, such as force against animals or accidental weapons discharges (both of which are removed during my data processing); based on guidance that officers are to use force to facilitate arrests or other law enforcement objectives when necessary to overcome a subject's resistance or to protect people or property (New Jersey Attorney General, 2000), force that does not result in an arrest should be rare and have a higher chance of being unjustified.

82% of all incidents in the force report have nonmissing charge information, with the median department's rate being 85%. However, this likely greatly overstates the actual frequency of force without charges due to the inability to differentiate between missing charges data and subjects who were not charged. The New Jersey State Police, for example, has 914 incidents in the cleaned force report data, making up more than 2% of all force reports. Of these 914 incidents, 911 have the charge field listed as "blank," one is "blnank," one is literally blank, and one is aggravated assault with a weapon. It is much more likely that the State Police systematically do not report charges on their force reports rather than virtually always using force against subjects who could not be charged with a crime. In contrast, the Atlantic City Police Department, which has the second-most force incidents in the data, has nonmissing charges for 96% of its incidents, and 50 out of the 455 departments are not missing charges in any of their force incidents. 38 departments with a combined 1,596 incidents are missing charges data in at least half of their reports, with 103 departments representing 4,417 incidents missing charges data in at least a third of them. There is also variation within departments suggestive of artificially missing charges. Newark, for example, has missing charges in 32% of its incidents, an unusually large amount for a department of its size, but this rate was less than 5% in 2012 before jumping between 30% and 51% in other years. Charge missingness is also not simply a result of officers using force against subjects not engaging in crime and therefore not being charged. If that were true, we would expect higher levels of force to be associated with lower levels of missing charges information, but this is not the case: incidents involving enhanced force are 20% more likely to have missing charge information than incidents with unenhanced force. Because larger departments are more likely to use higher levels of force and less likely to have missing charge information, I also check the force severity-missingness link for departments with at least 500 incidents to remove this source of variation. Even among these departments, unenhanced incidents here are missing charges 12% of the time, and enhanced incidents are missing charges 23% of the time. Taken together, the force reports appear to provide a lower bound of actual arrests in force incidents while suggesting a high rate of corresponding arrests, which may make arrests a useful benchmark for the set of incidents in which force might be used.

Appendix D. Department Leverage, Influence, and Identifying Variation

In regression models, an observation's *leverage* refers to how similar the values of its explanatory variables are to those of the rest of the sample; it is closely related to Mahalanobis distance. Given the same outcome, an observation with greater leverage, i.e., whose explanatory variable values are farther away from those of the others, will have a greater impact on the regression coefficients more than an observation with lower leverage will. An observation's *influence* measures how much it actually changes point estimates and can be computed by comparing coefficients between full and leave-one-out regressions that drop that observation.

Although computing each department's influence on the estimated racial disparities is simple—the group analog of a leave-one-out regression is simply dropping all observations from a department—the group notion of leverage is a recent innovation by MacKinnon, Nielsen, and Webb (2022) and encompasses not only regressor similarity but the size of the cluster. Their work also defines a cluster's *partial leverage*, its leverage with respect to a single coefficient of interest, such as the indicator for the subject being Black. Intuitively, it can be thought of as being similar to regression weights in that it determines how much potential a group has to exert actual influence on an outcome. Partial leverages sum to 1, so the average leverage is the inverse of the number of departments. Note that unlike influence, leverage does not depend on outcomes, and so there is only one vector of leverages, while influence will be different for each force outcome. With these tools, one can examine departments' partial leverages and actual influences on estimates, enabling diagnostics not unlike those popularized in the two-way fixed effects literature by Goodman-Bacon (2021). Given the number of departments and variation in numbers of incidents and numbers of Black subjects, this is especially valuable for understanding the disparity estimates.

The top-left graph of Figure A.9 shows a histogram of departmental partial leverages on the Black subject coefficient with a vertical line at the mean value. These values are based on the regression specification in Equation 1 but with year fixed effects instead of year-month to avoid problems of simultaneously identifying every parameter. The vast majority of departments have very low partial leverages, and the seven departments with partial leverages of at least 0.02 combine for more than a quarter of all partial leverage. The bottom-left graph of Figure A.9 plots the racial composition of adults 18-65 in each department's municipality against their partial leverage, again with a dashed line at the mean partial leverage value, which suggests that areas where Black residents make up about 20% to 50% of the population have the greatest partial leverages. However, it is clear from the final graph in the figure that the departments with the greatest leverage are those with the most force incidents, and these larger and/or higher-crime departments tend to have Black population proportions in that range.

Figure A.10 plots each department's influence, how much overall disparities from Equation 1 actually change at each outcome if a department is removed, against partial leverages for all force levels. Negative (positive) influence means that the overall disparity would be that much less (greater) without that department. This distribution is also skewed, with most departments exerting very little influence while a few are orders of magnitude more significant. But even those departments with the greatest influence do not by themselves have an appreciable impact on estimates given the size of the overall disparities (this is less true for the fired weapons outcome's small point estimate, where we should consider the comparatively wide confidence intervals). The relationship between partial leverage and influence magnitude is not particularly strong: it is a necessary condition for larger influence, but clearly not sufficient.

To present variation in the data visually, Figure A.11 plots geographic maps of the number of force incidents, population size, and the percentage of the population ages 18-65 that is Black for the municipalities with departments present in the data. There is a positive relationship between all three variables, but also exceptions, such as the high number of force incidents in relatively less-populous Camden and Atlantic City. See also the histograms in Figure A.2 of the distributions of each department's force incidents by race and per arrest.

In sum, there is variation in how much each department contributes to overall estimates of racial disparities that is very closely related to the number of force incidents for each department, but no one department has an appreciable impact on the aggregated racial disparity estimates.

Appendix E. Spatial Regressions

In this paper, I argue that individual departments are an important determinant of force usage. One key component in this claim is separating the effects of departments from those of space and geography. Are results due to the something about the departments, or are they based on of the composition of subjects and incidents within their borders? To investigate the importance of the departments, I leverage incident location information from the force reports to examine department impacts while holding geography fixed. The key to this strategy is that, in the course of policing, it is not uncommon for departments to use force in an area that would normally be the jurisdiction of another department. Manually cleaning the incident location data, I identify 1,350 incidents, about 3.4% of the data, where a department uses force within the borders of another municipality that has its own department.[19] Depending on how officers fill out the force reports, these could be incidents that start within the department's borders and end within another's, but this is still useful, with the logic being akin to that of a boundary discontinuity design: subjects within a small distance of the border are likely similar.

By examining how a department's force patterns change across different geographies, we can determine which is more predictive of a department's out-of-jurisdiction force usage: the department's typical force usage (supporting the importance of the department) or the force usage of the department that would normally be using force there (suggesting that geographic differences may be confounding departmental estimates).

My empirical strategy involves looping over each department that uses force within the borders of another department's municipality. For each such department, I use the union of two sets as the analytical sample. The first set is any incident where the department of interest uses force, regardless of location. The second is based on the set of other municipalities with police departments within which the department of interest uses force, taking all force incidents by the home department within their own borders. For example, suppose the New Brunswick Police Department is the department of interest. This department used force within their own borders as well as within the neighboring municipalities of Highland Park and East Brunswick. The estimation sample would then consist of all incidents involving the New Brunswick Police Department, regardless of location, and incidents from the Highland Park or East Brunswick police departments within their own borders.

For each department and corresponding data sample, I run regressions of the following form

$$Force_{iopt} = \beta \cdot \text{Black}_i + X'_{iopt}\gamma + \psi_{pj} + \nu_t + \varepsilon_{iopt}. \tag{E.1}$$

This equation is identical to Equation 1 except for the addition of the j subscript to note the department with jurisdiction in the corresponding municipality. Because of how the estimation sample is constructed, observations in which force was used by departments besides the department of interest (in the previous example, Highland Park and East Brunswick) will always have $p = j$. To visualize results, Figure A.12 plots coefficients ψ_{pj} with the New Jersey State Police as the depart-

[19] The actual incidence is higher, but New Jersey's myriad municipalities and unincorporated communities cannot always be uniquely identified by name, and I only mark an incident location as within another department's jurisdiction when there is no ambiguity. Warren County, for example, has both a Washington and a Washington Township, not to be confused with four other Washington Townships in the state or Robbinsville Township, which was known as Washington Township until 2008.

ment of interest. For reliability, I limit the locations to those where the State Police used force at least five times, and I focus on at least unenhanced force as the outcome, as it is used in half of all force incidents, giving maximum variation. Points are organized along the x-axis by the home department's regression coefficients ψ_{jj} (light points), allowing for comparison with the State Police's regression coefficients (dark points) in the same geographic area. It is clear that there is no relationship between the State Police's use of force in a municipality and the home department's use of force there ($\rho = -0.13$). Moreover, we can see that the State Police's use of force in most jurisdictions tends to be similar, showing a degree of consistency across locations (and officers).

To examine a broader sample, I now look at all departments who use force at least five times within their own borders (excepting the State Police) and five times within another department's jurisdiction. I compute the root mean squared error (RMSE) for two predictors of a department's out-of-jurisdiction force usage coefficients: the coefficients for the home departments, and the mean coefficient of the department's own force usage weighted by the number of incidents in each location. The department's own force usage outperforms the home department's force usage for 67% of these departments. Further, the average RMSE using the department's own force usage is 0.39, compared to 0.43 with the home department's coefficient as the predictor. Figure A.13 shows the relative performance of these predictors by plotting their respective RMSEs for each department. The results continue to suggest the importance of the department when holding fixed geography and that department fixed effects are reflective of the departments themselves, not simply their geography.

Appendix F. Predicting Department-Specific Disparities

What characteristics of the departments or their municipalities are associated with the department-specific racial disparities? For example, plotting a map of the disparities in Figure A.14 does not suggest geographic factors such as a municipality's proximity to New York City or Philadelphia, the nearest major metropolitan areas, or the presence of clusters of departments with larger racial differences. In this section, I treat the estimated department-specific disparities as the outcomes of interest to identify the factors that best predict racial inequities in police use of force. The goal of this exercise is to find suggestive evidence for possible contributors to racial inequities in police use of force that can serve as starting points for additional data collection or analysis to address why different police departments treat race differently. Although some of these observables may be manipulable by policymakers, this exercise does not directly inform policy: causality may run in many different directions, including between predictors, from outcome to predictor, or from unseen confounders.

For this prediction exercise, I use random forest models to handle many covariates relative to the number of departments. Under the framework of Gu and Walters (2022), a risk-neutral auditor deciding whether to undertake a costly investigation of a police department for inequitable practices would use a decision rule of whether the empirical Bayes posterior estimates are above a certain threshold. Based on this, I consider a binary classification problem analogous to a "horse race" regression of the outcome on a vector of departmental characteristics to see which variables are most informative.

$$\text{Top Decile}_p = X_p'\beta + \varepsilon_p \tag{F.1}$$

The positive outcome of interest is a department having a high disparity against Black subjects, which I define as being in the top decile for a given force level. The null outcome I use is a department having a weakly negative disparity, i.e., not having a disparity against Black subjects, and I omit from consideration departments in the middle, those with positive but less extreme disparities. When splitting the data into training and evaluation samples, I oversample observations with positive outcomes (using $2/3$ of them in the training sample, compared to $1/2$ of the null observations) to reduce class imbalance and the problems that may come with it, such as lazy algorithms that do not predict positive cases.

Because the mapping from police departments to municipalities is almost one-to-one, I include predictors based on both the police departments themselves and their municipalities. Doing so requires dropping the New Jersey State Police, as that department does not have a primary municipal jurisdiction. As predictors, I include a bevy of local characteristics that may be linked to racial policing: log median household income, Gini coefficient, log population, log population density, the percentage of the population ages 18-65 that is Black, violent crime rate, Mitt Romney's vote share in the 2012 presidential election, and county-level *D*-scores for the Black-White IAT from Project Implicit.[20] I then repeat this exercise on the non-random set of 92 departments in the 2016 LEMAS survey using additional measures of department policies and culture with plausible

[20] Refer to the introduction for factors that previous papers have found to be linked. Certain departmental characteristics that may be of interest, such as the number of officers and their race, are highly correlated with their municipal counterparts, population size and race. To avoid the difficulty of disentangling closely related variables with a moderate sample size and possible measurement error, I use only the municipal versions, as the municipal data from the American Community Survey are higher quality.

connections to racial disparities or general policing quality: engaging in asset forfeiture, using a written aptitude test in selecting recruits, having a written community policing plan, formally surveying local residents to improve policing, having written policies on stop and frisk, foot pursuits, and unbiased policing, and whether the department's chief executive is female or non-Hispanic White.[21] These variables may be manipulable by policymakers, but their limited coverage warrants extra caution in interpretation.

Figure A.15 presents the results for the first set of three random forest models, one for each outcome. The models for at least unenhanced force, enhanced force, and fired weapons have error rates of 46%, 27%, and 19%, respectively.[22] The graphs show the mean decrease classification accuracy associated with removing that predictor from the model. For the at least unenhanced force model, median household income is the only productive predictor. Enhanced force and firing weapons both have the Romney vote share as the most important predictor, and the violent crime rate is also the third best predictor for both.

The additional departmental characteristics from the 2016 LEMAS are not helpful for predicting departmental disparities. Figure A.16 plots the random forest results incorporating both the original variables and the new LEMAS variables. Among the three best predictors for each force level, only one of these nine, the use of a written aptitude test in selecting recruits in the at least unenhanced model, is from the LEMAS. The error rates for the three models have also increased slightly on average to 57%, 34%, and 18%. While these results do not suggest that these policy and cultural indicators are especially valuable predictors of racial disparities in police use of force, future investigations with a larger sample or different variables could have different results.

This prediction exercise yields several important takeaways about the factors that differentiate high- and no-disparity departments. First, predicting departmental disparities purely from characteristics of the departments and the municipalities they cover is difficult. This is also an area for improvements in centralized data collection like New Jersey has implemented, and better data availability across places and time can only improve our understanding of the most important departmental (or even officer) characteristics. Second, certain factors *are* useful predictors that can be investigated further to improve our understanding of where disparities come from. Third, the top predictors are not identical across force levels, which may suggest that not all racial disparities have the same underlying causes. And fourth, the most informative predictors tend to be characteristics of the municipality rather than the department. Overall, these findings underscore the variation and heterogeneity in disparities seen throughout this project and will hopefully serve as a starting point for future research to identify and ameliorate racial inequities in police use of force.

[21] Some variables of interest are inappropriate for this analysis due to insufficient variation. For example, all departments present in both my sample and the 2016 LEMAS allow at least some of their sworn officers with general arrest powers collective bargaining abilities.

[22] Because the training and evaluation data are not split identically between departments in the top decile of disparities and those without disparities, I compute the overall error rates as a weighted average of the type I and type II error rates to match the population proportions.

Table A.1: Summary Statistics for Police Departments Represented in Use of Force Data

Statistic	N	Mean	St. Dev.	Median	Min	Max
Avg. num. full-time police	455	42.75	76.41	23	0	1,088
% officers White	413	79.78	27.60	92.86	0.00	100.00
% officers Black	413	2.91	7.09	0.00	0.00	77.81
% officers Hispanic	413	2.24	5.94	0.00	0.00	53.49
% officers Asian/Pacific Islander	413	0.58	2.23	0.00	0.00	30.43
Total arrests 2012-2016	455	2,959.03	7,190.05	1,327	67	117,680
Black arrests 2012-2016	455	1,113.95	3,611.05	251	2	42,255
Force incidents 2012-2016	455	86.42	181.12	34	1	1,818
Black force incidents 2012-2016	455	35.13	108.01	6	0	1,172

Notes: Data cover all police departments in New Jersey from 2012 through 2016 with force reports after cleaning and processing. Not all departments' racial breakdowns sum to 100%. Data from ProPublica and NJ Advance Media.

Table A.2: Summary Statistics for Municipalities Represented in Use of Force Data

Statistic	N	Mean	St. Dev.	Median	Min	Max
Population	454	18,424.43	25,402.93	10,373	296	277,140
Population/square mile)	454	4,020.84	5,616.57	2,612.12	39.13	55,880.00
Median household income	454	87,296.63	31,652.80	83,006	26,214	190,625
Gini coefficient	454	0.43	0.05	0.43	0.33	0.60
Land area (sq. miles)	454	11.16	16.50	3.67	0.10	111.13
Pop. 18-65 % White	454	68.52	22.53	74.44	1.82	99.29
Pop. 18-65 % Black	454	8.41	12.17	3.87	0.00	88.77
Pop. 18-65 % Hispanic	454	14.16	14.30	9.19	0.00	82.94
Pop. 18-65 % Asian/PI	454	7.66	9.15	4.51	0.00	58.60
Violent crimes per 1000 residents	454	1.67	2.59	0.85	0.00	25.66
Romney vote share 2012 presidential election	454	45.20	14.79	47.74	1.31	81.82
County mean Black-White IAT *D*-score	21	0.35	0.05	0.35	0.24	0.41

Notes: Data cover all municipalities present in the force reports after cleaning and processing that are served by their own police department in the force reports data, i.e. not those served by New Jersey State Police. Data have been cleaned as described in Appendix A. Data come from the 2010 Census, 2012-2016 American Community Survey five-year estimates, FBI's Uniform Crime Reporting program, the New Jersey Division of Elections, and Project Implicit.

Table A.3: Departmental Regressions

	Prop. Subjects Black	Log(Incidents)	Force Incidents/Arrest
Prop. Pop. 18-65 Black	1.837^{***}	2.708^{***}	0.021
	(0.153)	(0.813)	(0.018)
Prop. Pop. 18-65 Black^2	-1.141^{***}	-4.259^{***}	-0.039^{*}
	(0.176)	(1.251)	(0.021)
Log(population)	0.013^{*}	0.790^{***}	0.001
	(0.007)	(0.041)	(0.001)
Log(population density)	0.003	0.091^{***}	0.003^{***}
	(0.007)	(0.035)	(0.001)
Violent crimes per 1000 residents	0.832	191.334^{***}	1.317^{***}
	(2.862)	(29.959)	(0.488)
R^2	0.566	0.673	0.111
Num. obs.	454	454	454

$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Notes: Table reports OLS estimates from regressions run at the departmental level using cleaned force data for New Jersey from 2012-2016. The regression specification is $Y_p = X_p + \varepsilon_p$, where a department's outcome variable Y (column names) are regressed on a vector X of each of explanatory variables (row names), with p denoting the department or its municipality.

Table A.4: Summary Statistics for Force Reports (Selection Subsample)

Statistic	N	Mean
Max force: compliance hold	8,050	0.32
Max force: unenhanced	8,050	0.43
Max force: enhanced (non-firearms)	8,050	0.24
Max force: fired weapon	8,050	0.01
Max subject action: resisted	8,050	0.00
Max subject action: physical threat/attack	8,050	0.91
Max subject action: blunt weapon threat/attack	8,050	0.02
Max subject action: knife threat/attack	8,050	0.03
Max subject action: vehicular threat/attack	8,050	0.02
Max subject action: firearm threat	8,050	0.02
Max subject action: fired weapon	8,050	0.003
Officer injured	8,050	0.17
Incident: crime in progress	8,050	0.47
Incident: domestic dispute	8,050	0.27
Incident: other dispute	8,050	0.24
Incident: suspicious person	8,050	0.00
Incident: traffic stop	8,050	0.08
Incident: other	8,050	0.00
Subject: White	8,050	0.49
Subject: Black	8,050	0.40
Subject: Hispanic	8,050	0.10
Subject: Asian/Pacific Islander	8,050	0.01
Subject: female	8,050	0.19
Subject: age	8,050	30.74

Notes: Data cover all police departments in New Jersey from 2012 through 2016 and include incidents where the following two conditions hold. 1) The subject at least physically threatened or attacked an officer or another person. 2) The incident's type is one where the subject's race was less likely to have influenced the officer's decision to engage with the subject: crimes in progress, disputes, and traffic stops at night. Data have been restructured so that each observation is a subject who had force used against them by police and cleaned as described in Appendix A. "Max force" refers to the highest level of force used in an incident; lower levels of force are not reported. "Max subject action" indicates the most severe action a subject took that could justify an office using force. Force used and subject actions are ordered from least severe to most severe. Incidents may have multiple types.

Table A.5: Summary Statistics for Municipalities with Departments Kept after Subsetting Use of Force Data

Statistic	N	Mean	St. Dev.	Median	Min	Max
Population	425	19,194.00	25,956.12	11,032	337	277,140
Population/square mile)	425	4,182.67	5,758.94	2,794.50	108.95	55,880.00
Median household income	425	85,637.58	31,071.57	80,950	26,214	190,625
Gini coefficient	425	0.43	0.05	0.43	0.33	0.60
Land area (sq. miles)	425	11.31	16.83	3.64	0.10	111.13
Pop. 18-65 % White	425	67.81	22.83	73.73	1.82	98.98
Pop. 18-65 % Black	425	8.70	12.46	4.12	0.00	88.77
Pop. 18-65 % Hispanic	425	14.64	14.62	9.37	0.00	82.94
Pop. 18-65 % Asian/PI	425	7.58	9.03	4.53	0.00	58.60
Violent crimes per 1000 residents	425	1.76	2.65	0.96	0.00	25.66
Romney vote share 2012 presidential election	425	44.48	14.77	46.86	1.31	73.25

Notes: Data cover the municipalities dropped from the cleaned data after limiting the data to incidents where the following two conditions hold. 1) The subject at least physically threatened or attacked an officer or another person. 2) The incident's type is one where the subject's race was less likely to have influenced the officer's decision to engage with the subject: crimes in progress, disputes, and traffic stops at night. Data come from the 2010 Census, 2012-2016 American Community Survey five-year estimates, FBI's Uniform Crime Reporting program, the New Jersey Division of Elections, and Project Implicit.

Table A.6: Summary Statistics for Municipalities with Departments Dropped by Subsetting Use of Force Data

Statistic	N	Mean	St. Dev.	Median	Min	Max
Population	29	7,146.34	9,869.39	4,160	296	53,238
Population/square mile)	29	1,649.18	1,387.58	1,587.79	39.13	5,918.79
Median household income	29	111,610.20	30,589.39	114,545	47,411	174,432
Gini coefficient	29	0.45	0.07	0.46	0.33	0.60
Land area (sq. miles)	29	8.84	10.61	4.80	0.28	45.23
Pop. 18-65 % White	29	78.84	14.20	82.11	45.15	99.29
Pop. 18-65 % Black	29	4.14	4.91	2.69	0.00	22.35
Pop. 18-65 % Hispanic	29	7.03	3.90	5.86	0.36	18.66
Pop. 18-65 % Asian/PI	29	8.80	10.87	3.28	0.00	33.17
Violent crimes per 1000 residents	29	0.35	0.63	0.17	0.00	3.34
Romney vote share 2012 presidential election	29	55.87	10.43	53.90	32.71	81.82

Notes: Data cover the municipalities present in both the original cleaned data and the subsetted data from limiting the sample to incidents where the following two conditions hold. 1) The subject at least physically threatened or attacked an officer or another person. 2) The incident's type is one where the subject's race was less likely to have influenced the officer's decision to engage with the subject: crimes in progress, disputes, and traffic stops at night. Data come from the 2010 Census, 2012-2016 American Community Survey five-year estimates, FBI's Uniform Crime Reporting program, the New Jersey Division of Elections, and Project Implicit.

Table A.7: Summary Statistics for Police Departments Kept after Subsetting Use of Force Data

Statistic	N	Mean	St. Dev.	Median	Min	Max
Avg. num. full-time police	426	44.56	78.53	24	0	1,088
% officers White	392	79.16	27.94	92.01	0.00	100.00
% officers Black	392	2.97	7.08	0.00	0.00	77.81
% officers Hispanic	392	2.28	5.86	0.00	0.00	53.49
% officers Asian/Pacific Islander	392	0.61	2.28	0.00	0.00	30.43
Total arrests 2012-2016	426	3,127.13	7,400.54	1,425.5	99	117,680
Black arrests 2012-2016	426	1,183.92	3,721.80	284	3	42,255
Force incidents 2012-2016	426	91.86	185.92	37	1	1,818
Black force incidents 2012-2016	426	37.45	111.26	6	0	1,172

Notes: Data cover all police departments in New Jersey from 2012 through 2016 with force reports after cleaning and processing. Not all departments' racial breakdowns sum to 100%. Data from ProPublica and NJ Advance Media.

Table A.8: Summary Statistics for Police Departments Dropped by Subsetting Use of Force Data

Statistic	N	Mean	St. Dev.	Median	Min	Max
Avg. num. full-time police	29	16.21	17.21	12	4	96
% officers White	21	91.51	16.70	100.00	50.00	100.00
% officers Black	21	1.95	7.38	0.00	0.00	33.33
% officers Hispanic	21	1.59	7.27	0.00	0.00	33.33
% officers Asian/Pacific Islander	21	0.09	0.42	0.00	0.00	1.92
Total arrests 2012-2016	29	489.72	425.29	339	67	1,915
Black arrests 2012-2016	29	86.10	99.21	43	2	436
Force incidents 2012-2016	29	6.48	13.65	3	1	74
Black force incidents 2012-2016	29	1.00	1.83	0	0	9

Notes: Data cover all police departments in New Jersey from 2012 through 2016 with force reports after cleaning and processing. Not all departments' racial breakdowns sum to 100%. Data from ProPublica and NJ Advance Media.

Table A.9: Racial Disparities by Police in Force of At Least Specified Severity, Conditional on Force (No Controls)

	Unenhanced	Enhanced	Fired Weapons
Subject Black	0.082^{***}	0.046^{***}	0.002^{*}
	(0.011)	(0.011)	(0.001)
Clustering	Dept.	Dept.	Dept.
Outcome mean	0.505	0.146	0.004
R^2	0.007	0.004	0.000
Num. obs.	39321	39321	39321

$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Notes: Table reports OLS estimates from regressions where the outcome is a binary measure of whether police used force of at least the specified severity on indicators for subject race with no other covariates.

Table A.10: Racial Disparities by Police in Force of At Least Specified Severity, Conditional on Force

	Unenhanced		Enhanced		Fired Weapons	
Subject Black	0.018^{***}	0.016^{*}	0.034^{***}	0.037^{***}	0.001	0.001
	(0.006)	(0.009)	(0.005)	(0.006)	(0.001)	(0.001)
Subject Hispanic	−0.001	0.005	0.005	0.007	0.001	0.001
	(0.010)	(0.013)	(0.007)	(0.010)	(0.002)	(0.002)
Subject Asian/PI	−0.024	−0.008	0.005	0.017	-0.005^{***}	−0.002
	(0.025)	(0.040)	(0.017)	(0.026)	(0.002)	(0.002)
Fixed effects	Dept.	Off.	Dept.	Off.	Dept.	Off.
Clustering	Dept.	Off.	Dept.	Off.	Dept.	Off.
Outcome mean	0.505	0.505	0.146	0.145	0.004	0.004
R^2	0.182	0.512	0.118	0.464	0.182	0.532
Num. obs.	39183	39130	39183	39130	39183	39130

$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Notes: Table reports OLS estimates from Equation 1, which regresses a binary measure of whether police used force of at least the specified severity on indicators for the subject's race, an indicator for the subject being female, a quadratic of the subject's age, time of day fixed effects, indicators for the subject's actions, an indicator for whether an officer was harmed, an indicator for whether the subject was an emotionally disturbed person, indicators for the officer's rank, year fixed effects, and department or officer fixed effects.

Table A.11: Conditional Logit Odds Ratios on Race in Intensity of Force Used, Conditional on Force

	Unenhanced	Enhanced	Fired Weapons
Subject Black	1.054^{***}	1.286^{***}	1.090
	$(1.024, 1.085)$	$(1.203, 1.374)$	$(0.722, 1.646)$
Fixed effects	Dept.	Dept.	Dept.
Clustering	Dept.	Dept.	Dept.
Num. obs.	39321	39321	39321

$^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Notes: Table reports conditional logit estimates from Equation 2, which regresses a binary measure of whether police used force of at least the specified severity on indicators for the subject's race, an indicator for the subject being female, a quadratic of the subject's age, time of day fixed effects, indicators for the subject's actions, an indicator for whether an officer was harmed, an indicator for whether the subject was an emotionally disturbed person, indicators for the officer's rank, and year fixed effects, stratified by department. Asymmetric 95% confidence intervals with clustering at the department level based on exponentiating the log odds confidence interval are in parentheses.

Table A.12: Ordered Logit Odds Ratios on Race in Policing Outcomes

	Intensive Margin	Arrests+Intensive Margin
Subject Black	1.148^{***}	1.303^{***}
	$(1.102, 1.197)$	$(1.270, 1.338)$
Sample	Force Microdata	Aggregated Force+Arrests
Num. obs.	39183	899921

$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Notes: Table reports ordered logit estimates. The first column uses the same data and regressors as the main regressions from Equation efeq:main, regressing the level of force used by police in an incident on indicators for the subject's race, an indicator for the subject being female, a quadratic of the subject's age, time of day fixed effects, indicators for the subject's actions, an indicator for whether an officer was harmed, an indicator for whether the subject was an emotionally disturbed person, indicators for the officer's rank, and year fixed effects, stratified by department. The model in the second column supplements the force reports with aggregated arrests data from the FBI's Uniform Crime Reporting program, treating arrests as the lowest outcome in an incident. The number of arrests-only incidents is calculated as the reported number of arrests less the number of force incidents, imputing values of 0 for the 1.7% percent of race-age-department-month cells with negative differences. Because of the lack of microdata from the FBI data, grouped observations are weighted by the number of incidents covered, and the regressors are indicators for the subject being Black, the subject being at least 18 years old, the incident's year-month, and the department. Asymmetric 95% confidence intervals based on exponentiating the log odds confidence interval are in parentheses.

Table A.13: Racial Disparities by Police in Force of At Least Specified Severity, Conditional on Force, Incidents where Race Likely Unrelated to Decision to Engage with Subject and Subject at Least Physically Threatened/Attacked Officer or Another

	Unenhanced		Enhanced		Fired Weapons	
Subject Black	0.041***	0.024	0.058***	0.045*	0.002	−0.001
	(0.014)	(0.028)	(0.014)	(0.027)	(0.002)	(0.004)
Fixed effects	Dept.	Off.	Dept.	Off.	Dept.	Off.
Clustering	Dept.	Off.	Dept.	Off.	Dept.	Off.
Outcome mean	0.679	0.679	0.247	0.247	0.009	0.009
R^2	0.191	0.722	0.137	0.700	0.279	0.839
Num. obs.	8029	8019	8029	8019	8029	8019

$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Notes: Table reports OLS estimates from Equation 1, which regresses a binary measure of whether police used force of at least the specified severity on indicators for the subject's race, an indicator for the subject being female, a quadratic of the subject's age, time of day fixed effects, indicators for the subject's actions, an indicator for whether an officer was harmed, an indicator for whether the subject was an emotionally disturbed person, indicators for the officer's rank, year fixed effects, and department or officer fixed effects. These regressions are run on a subset of the data where the following two conditions hold. 1) The subject at least physically threatened or attacked an officer or another person. 2) The incident's type is one where the subject's race was less likely to have influenced the officer's decision to engage with the subject: crimes in progress, disputes, and traffic stops at night.

Table A.14: Racial Disparities in Number of Officers Using Force in Incident

	1	2
Subject Black	0.078^{***}	0.093^{***}
	(0.014)	(0.017)
Fixed effects	Dept.	Off.
Clustering	Dept.	Off.
R^2	0.117	0.483
Num. obs.	34920	34872

$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Notes: Table reports OLS estimates based on a modified version of Equation 1 where the outcome is the number of officers who used force in that incident. This is regressed on an indicator for the subject being Black, an indicator for the subject being female, a quadratic of the subject's age, time of day fixed effects, indicators for the subject's actions, an indicator for whether an officer was harmed, an indicator for whether the subject was an emotionally disturbed person, indicators for the officer's rank, year fixed effects, and department or officer fixed effects.

Table A.15: Racial Disparities by Police in Force of At Least Specified Severity, Conditional on Force, with Number of Officers Using Force in Incident

	Unenhanced		Enhanced		Fired Weapons	
Subject Black	0.019**	0.012	0.036***	0.036***	0.002	0.001
	(0.008)	(0.010)	(0.006)	(0.008)	(0.001)	(0.001)
2 Officers Using Force	0.033***	0.044***	0.019***	0.026***	−0.001	−0.000
	(0.007)	(0.010)	(0.006)	(0.007)	(0.001)	(0.001)
3 Officers Using Force	0.107***	0.123***	0.065***	0.080***	−0.001	−0.000
	(0.011)	(0.015)	(0.009)	(0.012)	(0.001)	(0.002)
4+ Officers Using Force	0.158***	0.168***	0.110***	0.115***	−0.000	0.001
	(0.015)	(0.020)	(0.012)	(0.018)	(0.002)	(0.003)
Subject Black × 2 Officers Using Force	0.002	0.013	−0.019**	−0.017	−0.002	−0.000
	(0.011)	(0.015)	(0.008)	(0.012)	(0.002)	(0.002)
Subject Black × 3 Officers Using Force	−0.013	−0.013	0.008	0.007	−0.000	−0.001
	(0.018)	(0.023)	(0.012)	(0.019)	(0.002)	(0.003)
Subject Black × 4+ Officers Using Force	−0.028	−0.031	−0.001	0.017	−0.005*	−0.003
	(0.023)	(0.029)	(0.018)	(0.026)	(0.002)	(0.003)
Fixed effects	Dept.	Off.	Dept.	Off.	Dept.	Off.
Clustering	Dept.	Off.	Dept.	Off.	Dept.	Off.
Outcome mean	0.487	0.505	0.139	0.145	0.004	0.004
R^2	0.189	0.520	0.125	0.472	0.179	0.524
Num. obs.	38600	38548	38600	38548	38600	38548

$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Notes: Table reports OLS estimates from Equation 1, which regresses a binary measure of whether police used force of at least the specified severity on indicators for the subject's race, an indicator for the subject being female, a quadratic of the subject's age, time of day fixed effects, indicators for the subject's actions, an indicator for whether an officer was harmed, an indicator for whether the subject was an emotionally disturbed person, indicators for the officer's rank, year fixed effects, and department or officer fixed effects. These regressions additionally include indicators for the number of unique officers who use force in a given incident across all subjects and its interaction with the subject being Black. To avoid improper comparisons, the sample includes only incidents where all or no subjects are Black.

Table A.16: Racial Disparities by Police in Force of At Least Specified Severity, Conditional on Force, with Officer Experience

	Unenhanced	Enhanced	Fired Weapons
Subject Black	0.024^{**}	0.028^{***}	-0.002
	(0.012)	(0.008)	(0.001)
Log(Years Officer Experience)	0.038^{***}	0.026^{***}	0.001
	(0.005)	(0.003)	(0.001)
Subject Black × Log(Years Officer Experience)	-0.001	0.006	0.001
	(0.006)	(0.004)	(0.001)
Fixed effects	Dept.	Dept.	Dept.
Clustering	Dept.	Dept.	Dept.
Outcome mean	0.487	0.139	0.004
R^2	0.188	0.136	0.174
Num. obs.	27317	27317	27317

$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Notes: Table reports OLS estimates from Equation 1, which regresses a binary measure of whether police used force of at least the specified severity on indicators for the subject's race, an indicator for the subject being female, a quadratic of the subject's age, time of day fixed effects, indicators for the subject's actions, an indicator for whether an officer was harmed, an indicator for whether the subject was an emotionally disturbed person, indicators for the officer's rank, year fixed effects, and department or officer fixed effects. These regressions additionally include the indicated measures of the officer's experience. The officer experience field is not always present, resulting in a smaller sample size; 89.9% of departments have at least some values, with the median department having 98.1% coverage.

_____________________ POLICE DEPARTMENT

USE OF FORCE REPORT

A. Incident Information

Date	Time	Day of Week	Location	**INCIDENT NUMBER**

Type of Incident
☐ Crime in progress ☐ Domestic ☐ Other dispute ☐ Suspicious person ☐ Traffic stop
☐ Other (specify)

B. Officer Information

Name (Last, First, Middle)		Badge #	Sex	Race	Age	Injured Y / N	Killed Y / N
Rank	Duty assignment	Years of service		On-Duty Y / N		Uniform Y / N	

C1. Subject 1 (List only the person who was the subject of the use of force by the officer listed in Section B.)

Name (Last, First, Middle)	Sex	Race	Age	Weapon Y / N	Injured Y / N	Killed Y / N
☐ Under the influence ☐ Other unusual condition (specify)	Arrested Y / N		Charges			

Subject's actions (check all that apply)
☐ Resisted police officer control
☐ Physical threat/attack on officer or another
☐ Threatened/attacked officer or another with blunt object
☐ Threatened/attacked officer or another with knife/cutting object
☐ Threatened/attacked officer or another with motor vehicle
☐ Threatened officer or another with firearm
☐ Fired at officer or another
☐ Other (specify)

Officer's use of force toward this subject (check all that apply)
☐ Compliance hold
☐ Hands/fists
☐ Kicks/feet
☐ Chemical/natural agent
☐ Strike/use baton or other object
☐ Canine
☐ Other (specify)

Firearms Discharge
☐ Intentional
☐ Accidental

Number of Shots Fired _____
Number of Hits _____
[Use 'UNK' if unknown]

C2. Subject 2 (List only the person who was the subject of the use of force by the officer listed in Section B.)

Name (Last, First, Middle)	Sex	Race	Age	Weapon Y / N	Injured Y / N	Killed Y / N
☐ Under the influence ☐ Other unusual condition (specify)	Arrested Y / N		Charges			

Subject's actions (check all that apply)
☐ Resisted police officer control
☐ Physical threat/attack on officer or another
☐ Threatened/attacked officer or another with blunt object
☐ Threatened/attacked officer or another with knife/cutting object
☐ Threatened/attacked officer or another with motor vehicle
☐ Threatened officer or another with firearm
☐ Fired at officer or another
☐ Other (specify)

Officer's use of force toward this subject (check all that apply)
☐ Compliance hold
☐ Hands/fists
☐ Kicks/feet
☐ Chemical/natural agent
☐ Strike/use baton or other object
☐ Canine
☐ Other (specify)

Firearms Discharge
☐ Intentional
☐ Accidental

Number of Shots Fired _____
Number of Hits _____
[Use 'UNK' if unknown]

➤ **If this officer used force against more than two subjects in this incident, attach additional USE OF FORCE REPORTS.**

Signature:	Date:
Print Supervisor Name:	Supervisor Signature:

7/2001

Figure A.1: New Jersey Model Use of Force Report

Notes: Figure obtained from the website of the New Jersey Attorney General. Not all departments follow this template.

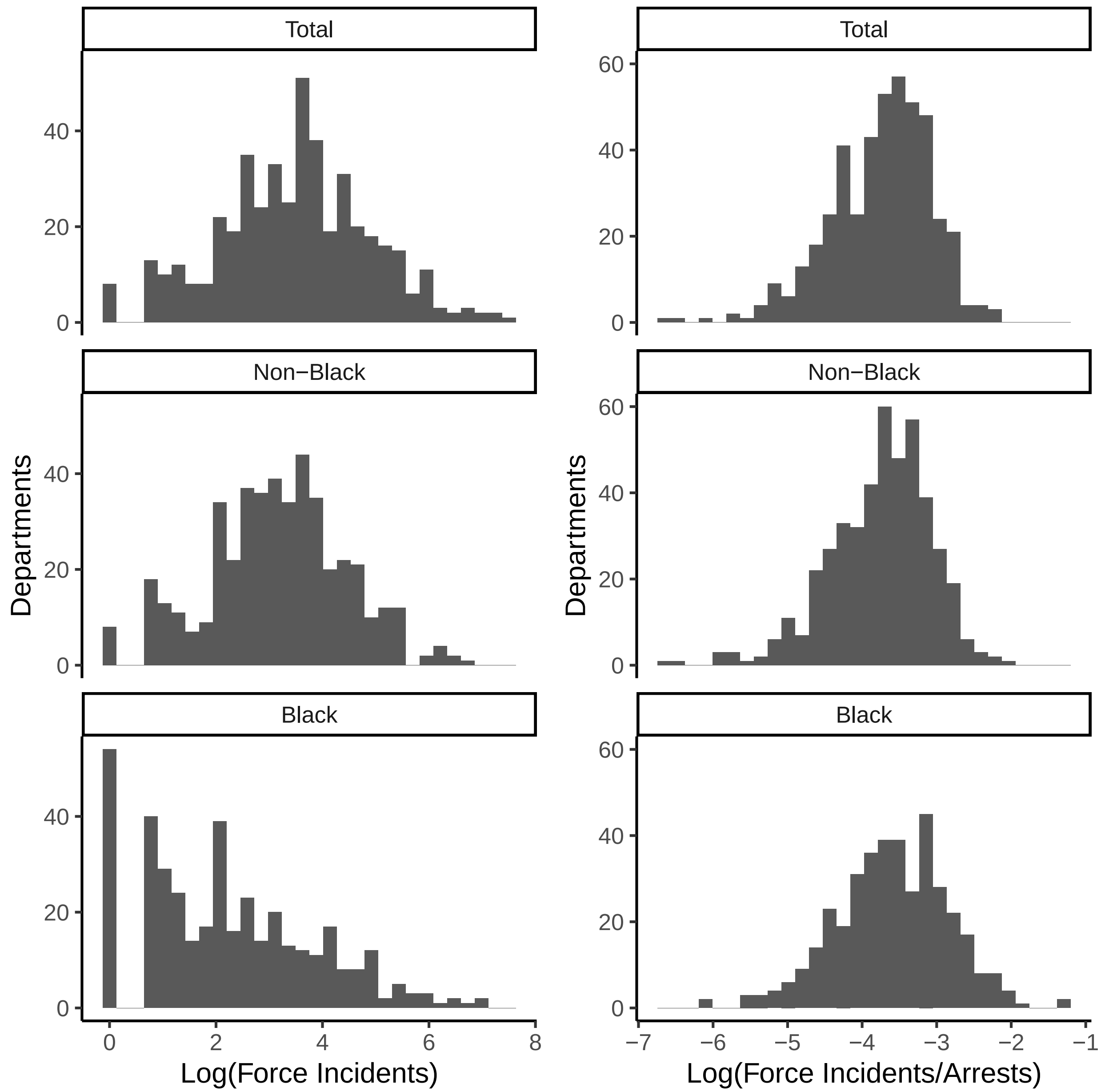

Figure A.2: Departmental Force Incidents and Force Incidents Per Arrest

Notes: Figure displays summary statistics for the departments present in the cleaned use of force data from New Jersey for 2012-2016. Graph titles refer to the race of the subject.

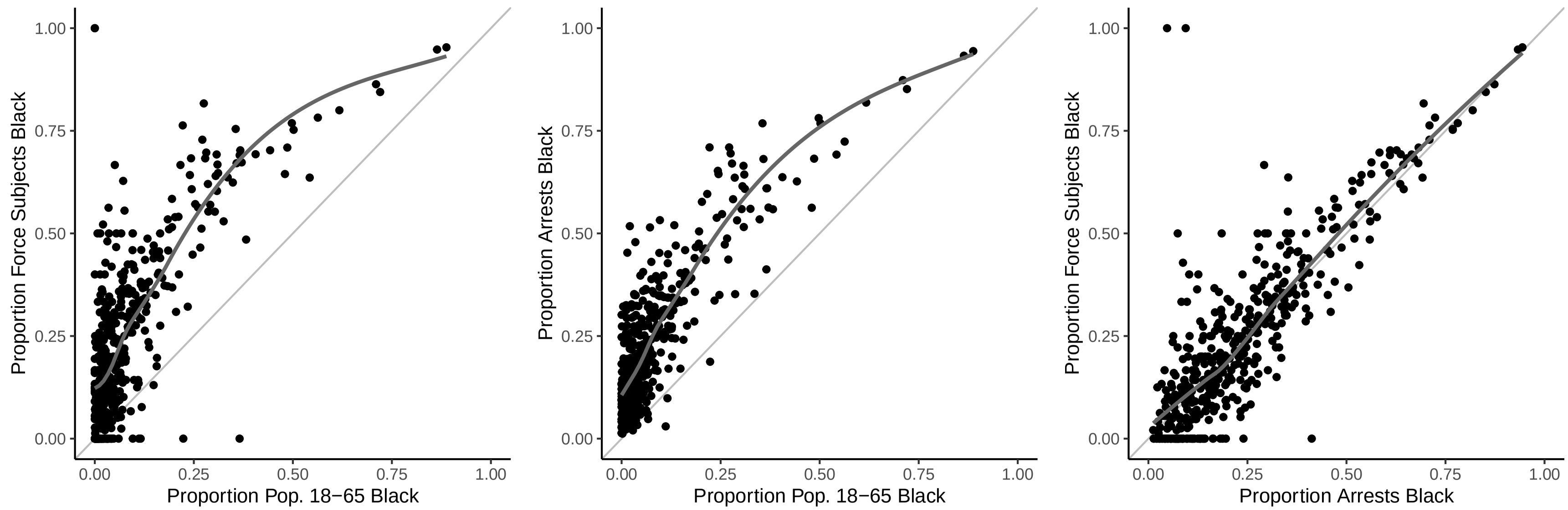

Figure A.3: Racial Composition of Local Population, Arrestees, and Force Subjects

Notes: Figure plots relationships between the Black proportion of the local population ages 18-65 and the Black proportions of force subjects and arrestees for the departments present in the cleaned use of force data from New Jersey for 2012-2016, except for the New Jersey State Police. Fit lines are calculated by loess.

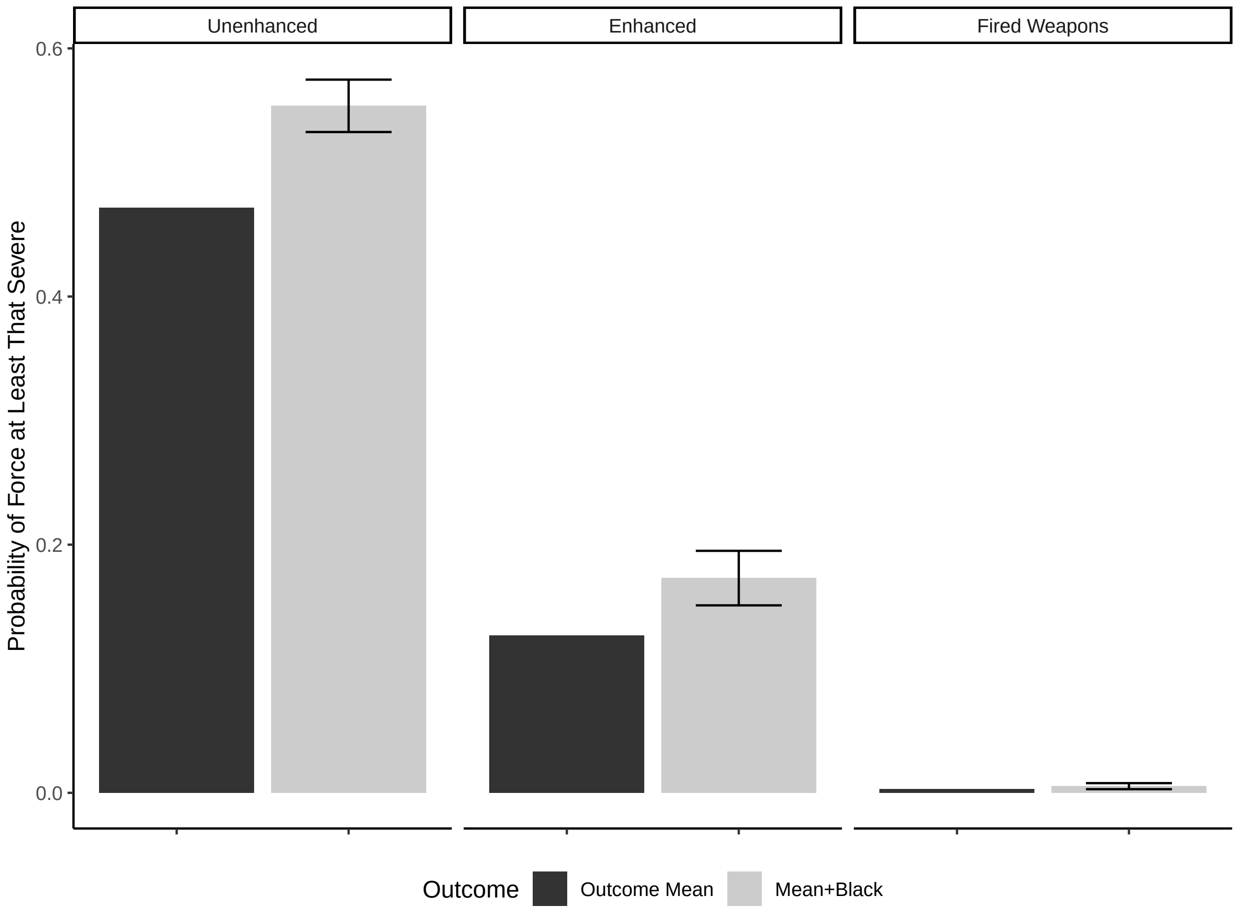

Figure A.4: Overall Racial Disparities (No Controls)

Notes: Figure presents results from a series of OLS models regressing outcomes on an indicator for the subject being Black. Each heading represents a different outcome: whether, conditional on any force being used, force of at least the specified severity was used. Black values are obtained by taking the outcome mean and adding the coefficient on being Black, with 95% confidence intervals based on the corresponding standard errors.

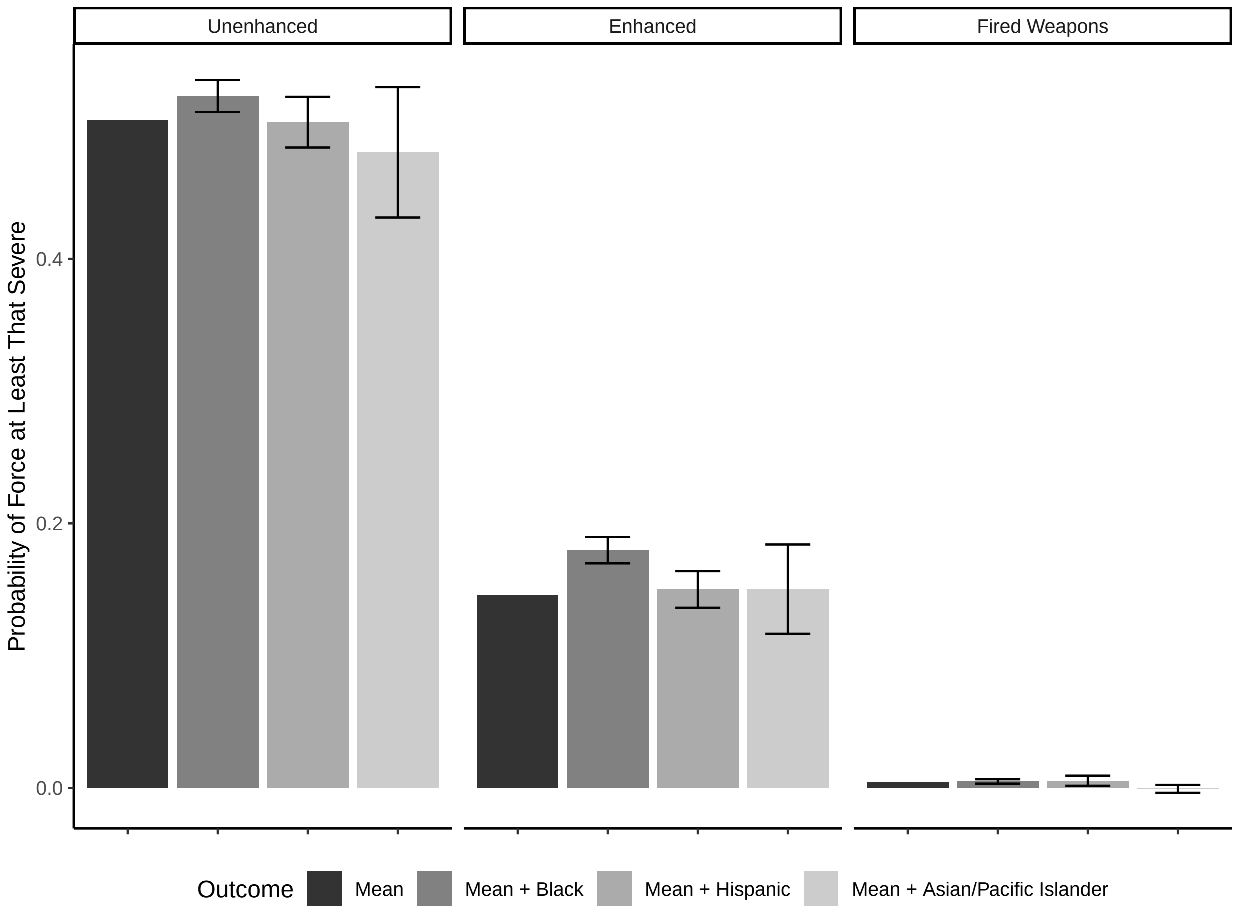

Figure A.5: Overall Racial Disparities of Subject Being Black on Probability of Force of at Least Specified Severity, Full Race Dummies (Department Fixed Effects and Clustering)

Notes: Figure presents results from a series of OLS models fit via Equation 1. Each heading represents a different outcome: whether, conditional on any force being used, force of at least the specified severity was used. 95% confidence intervals are based on the corresponding race coefficient. Standard errors are clustered at the department level.

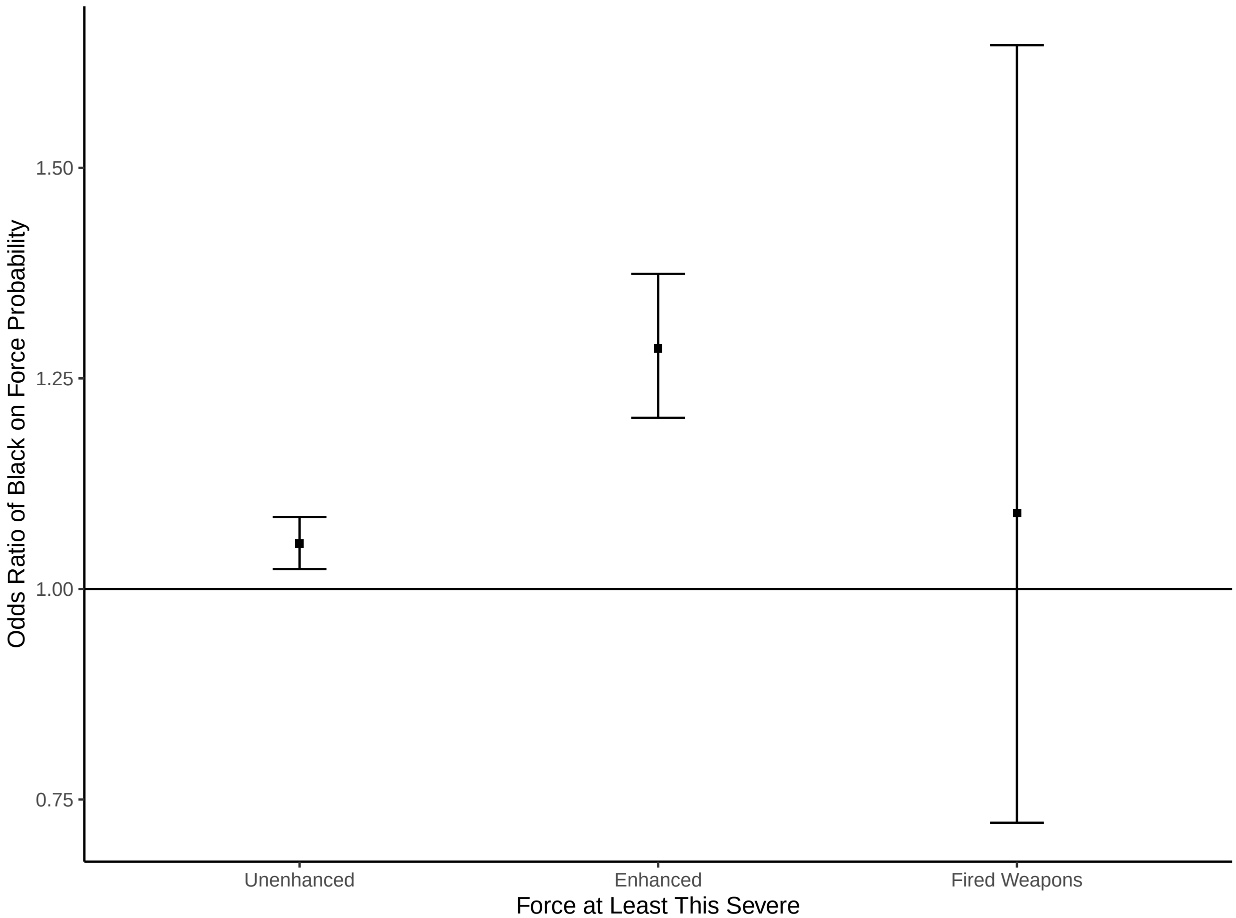

Figure A.6: Conditional Logit Odds Ratios of Subject Being Black on Probability of Force of at Least Specified Severity

Notes: Figure presents results from a series of conditional logit models fit via Equation 2. Each heading represents a different outcome: whether, conditional on any force being used, force of at least the specified severity was used. Points indicate the odds ratio of a Black subject. 95% confidence intervals clustered at the department level are obtained by exponentiating the logit coefficients from the regression and are asymmetric.

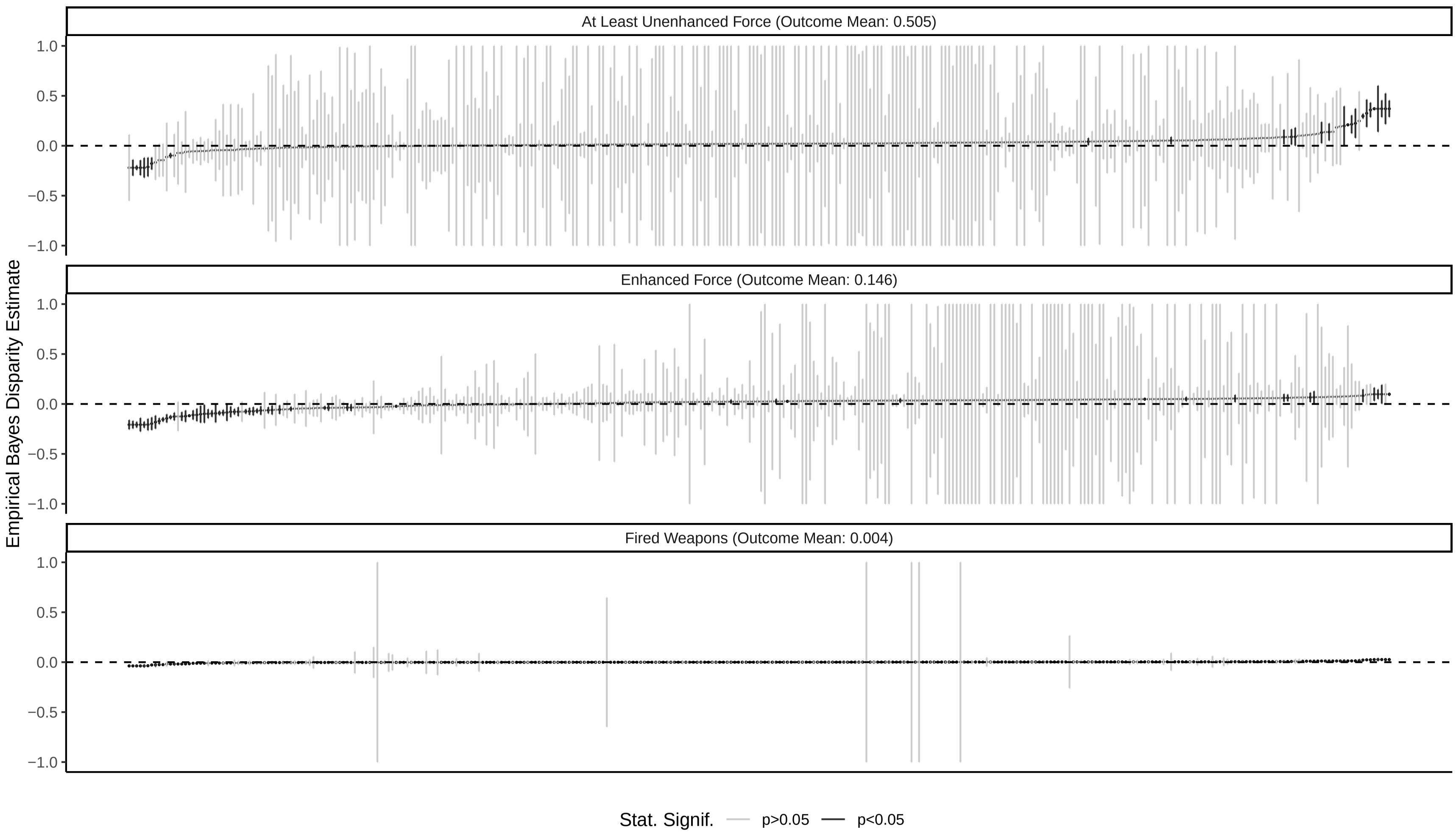

Figure A.7: Parametric 95% Confidence Intervals for Empirical Bayes Estimates of Departmental Disparities

Notes: Figure presents parametric 95% confidence intervals for empirical Bayes estimates of departmental racial disparities by force level using the procedure outlined in Morris (1983). Confidence interval bounds that extend beyond $[-1, 1]$ have been truncated, and fully shrunk estimates (i.e., those that place all weight on the grand mean and none on the point estimate) have been omitted.

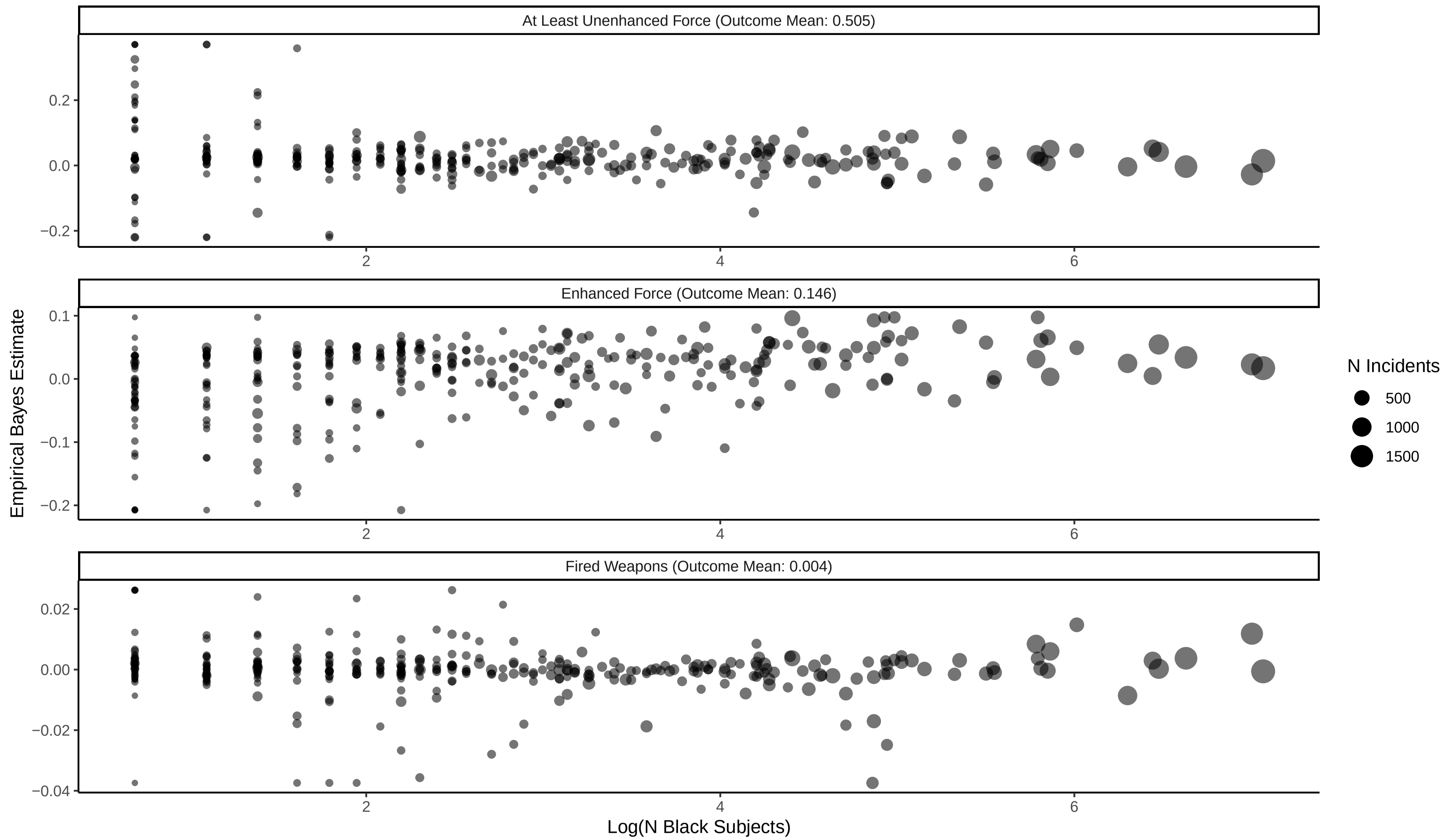

Figure A.8: Empirical Bayes Estimates of Departmental Disparities, Number of Incidents with Black Subjects, and Number of Incidents

Notes: Figure presents plots departmental empirical Bayes estimates of departmental racial disparities by force level against the number of incidents involving Black subjects, with point size corresponding to the number of incidents for that department.

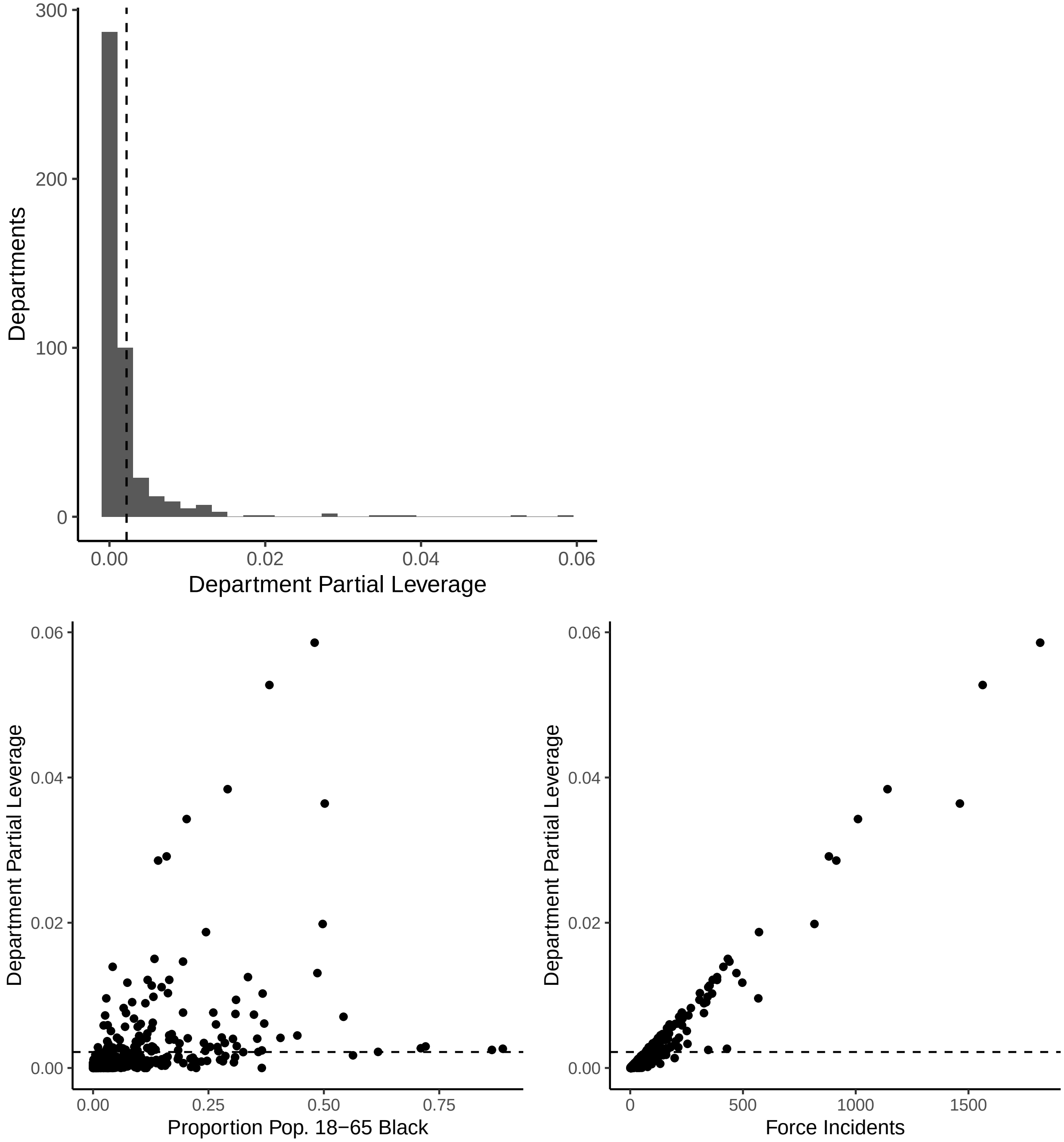

Figure A.9: Department Partial Leverages

Notes: Figure presents departmental partial leverages for the coefficient on the subject being Black as defined by MacKinnon, Nielsen, and Webb (2022). The corresponding regression specification is Equation 1 using year fixed effects instead of year-month to avoid simultaneous identification problems. Dashed lines indicate the average partial leverage value.

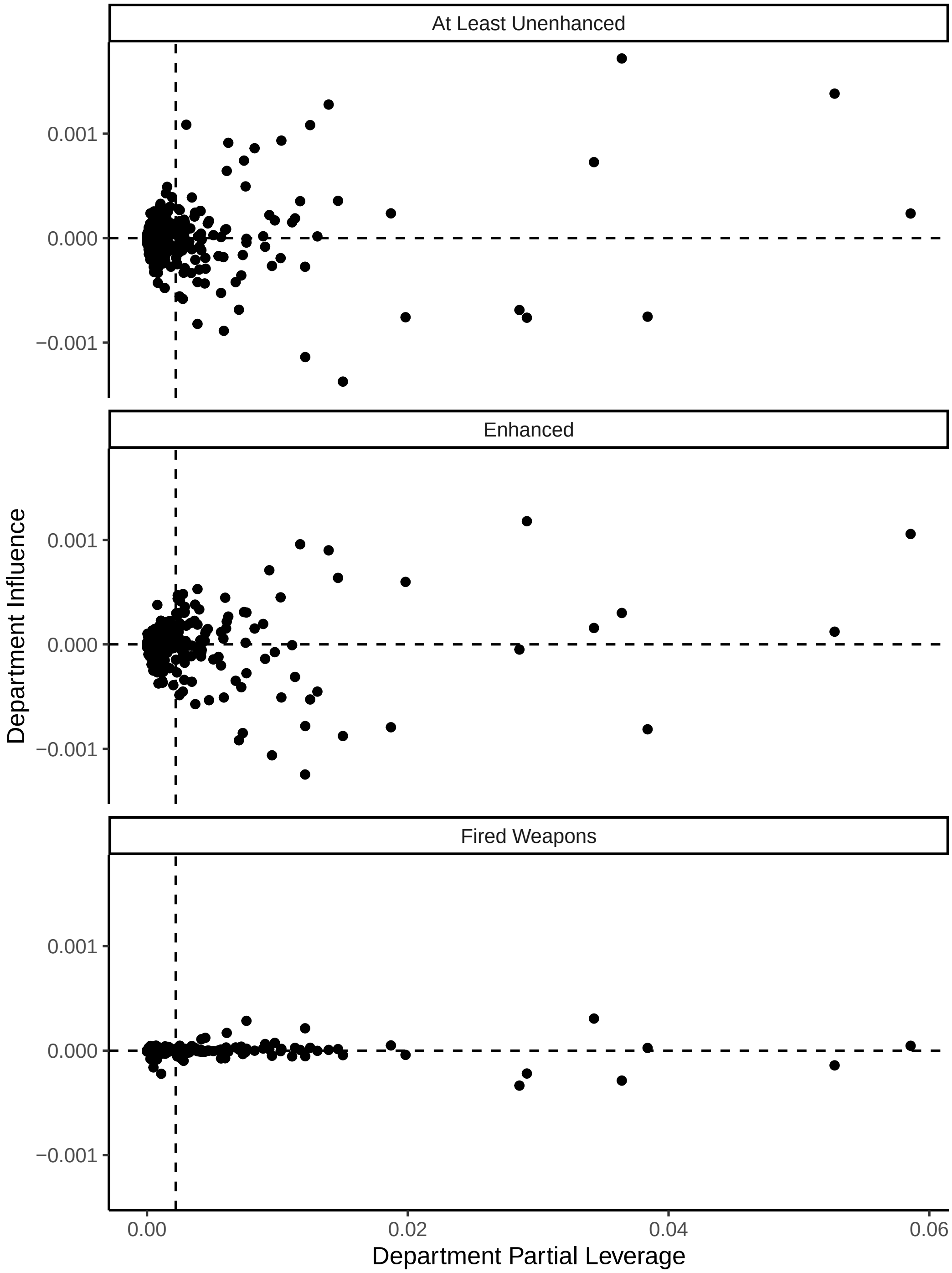

Figure A.10: Department Influences and Partial Leverages

Notes: Figure presents department influences against their partial leverages for the coefficient on the subject being Black as defined by MacKinnon, Nielsen, and Webb (2022). The regression specification is Equation 1, with the partial leverages using year fixed effects instead of year-month to avoid simultaneous identification problems. Positive (negative) influence indicates that the regression coefficient would have been that much greater (less) had the department not been included in the regression. Dashed lines indicate mean values.

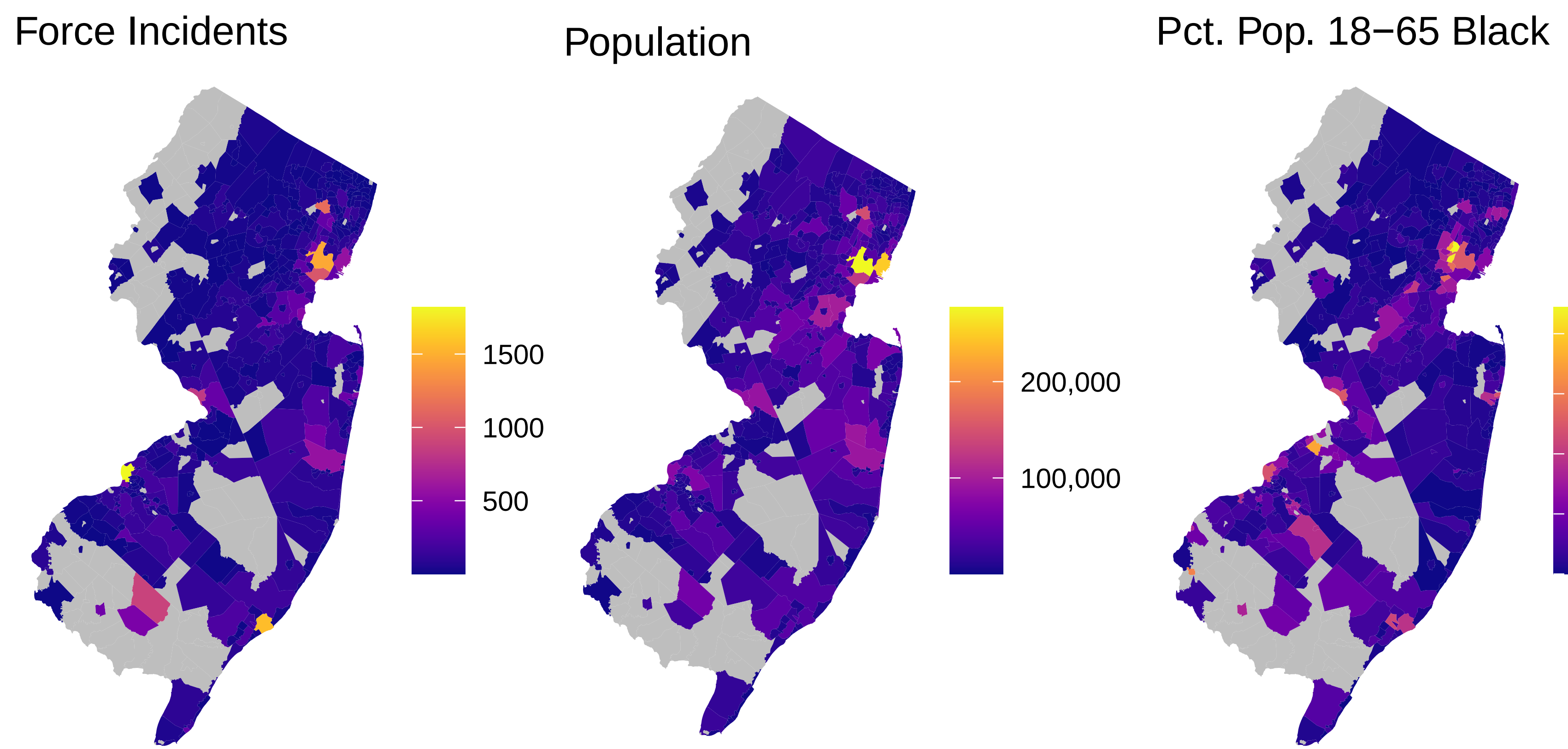

Figure A.11: Heatmap of Municipal Summary Statistics

Notes: Figure presents heatmaps of municipal summary statistics. Municipalities in gray do not have departments in the data.

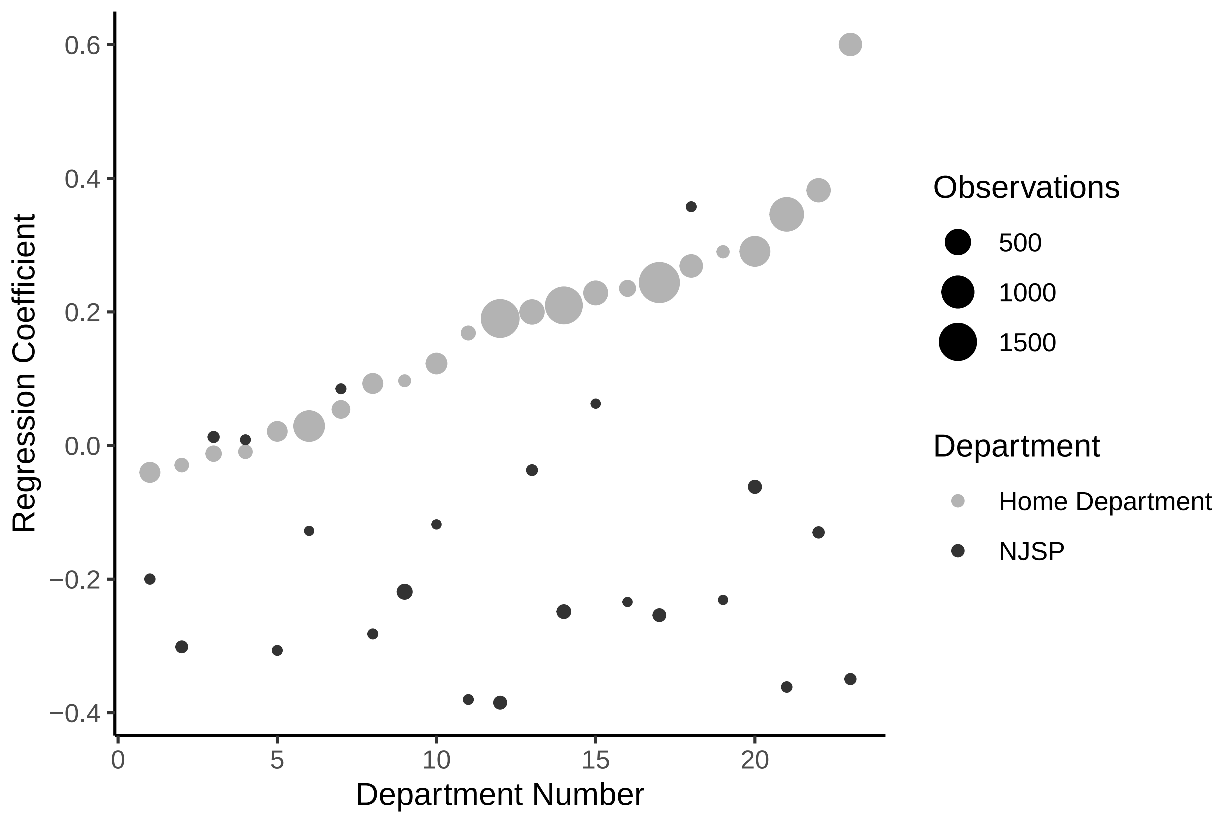

Figure A.12: Regression Coefficients for At Least Unenhanced Force, New Jersey State Police and Municipal Departments within Same Municipalities Indicators

Notes: Figure presents regression coefficients from Equation E.1 for the New Jersey State Police and the home department in municipalities where the NJSP used force at least five times. Each point is a department's coefficient for using at least unenhanced force in an area, i.e., the department-location fixed effect.

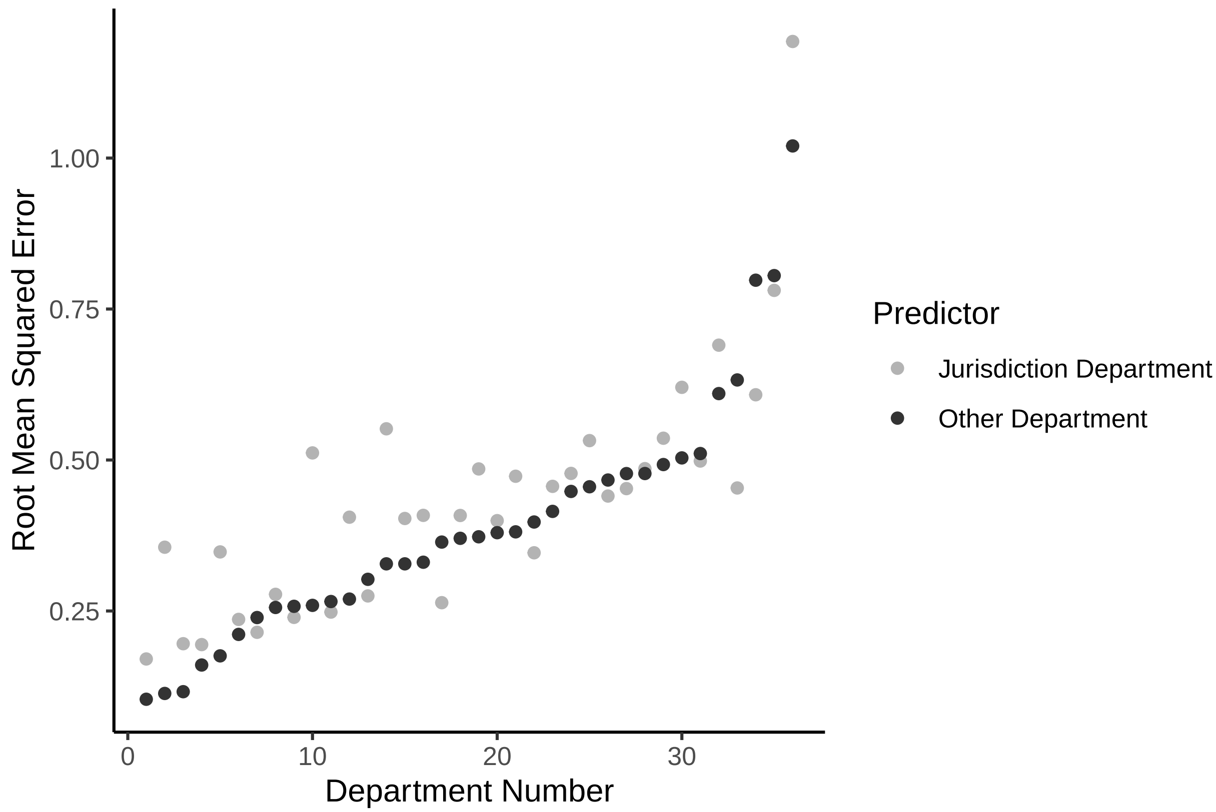

Figure A.13: Root Mean Squared Error of Different Predictors of Departments' Use of At Least Unenhanced Force within Another Department's Borders

Notes: Figure presents root mean squared errors of predictors of a department's force usage within other departments' jurisdictions as described by Equation E.1 for departments who use force at least five times within other departments' municipal borders. The light gray dots indicate the RMSE when predicting force coefficients using the home department's regression coefficient, while the dark dots are the RMSE when predicting the out-of-jurisdiction department's force usage using the weighted average of their force usage across locations.

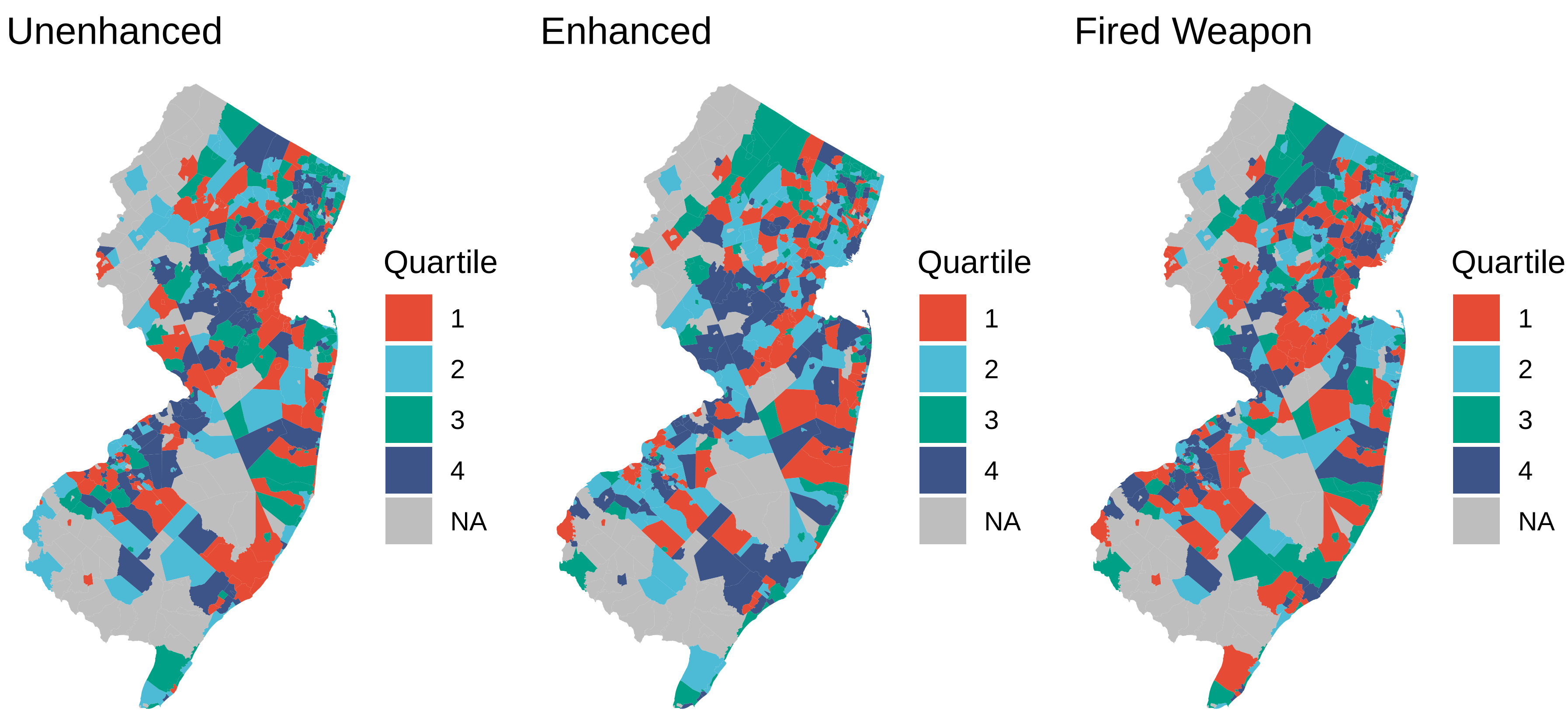

Figure A.14: Heatmap of Empirical Bayes Estimates of Department-Black Interactions

Notes: Figure presents heatmaps of department-specific racial disparities β between White/Asian/Pacific Islander subjects and Black subjects as estimated from the empirical Bayes estimator in Equation 5. Each subgraph shows results from regressions with the specified outcome outcome: whether, conditional on any force being used, force of at least the specified severity was used. Municipalities in gray do not have departments in the data.

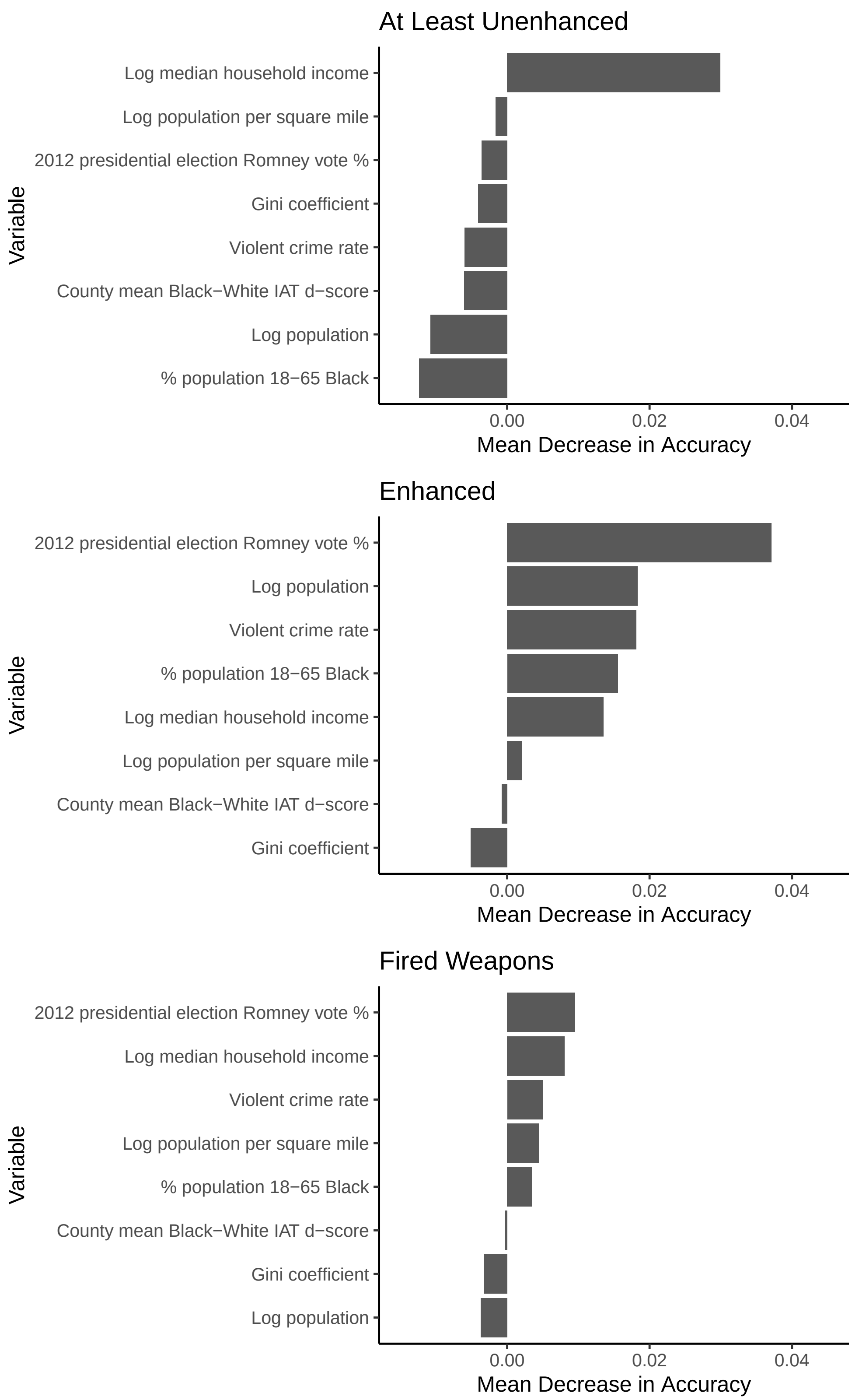

Figure A.15: Random Forest Variable Importance Plots

Notes: Figure presents variables' average effects on the accuracy of random forest models on classification models for whether a department has a disparity in its use of force against Black subjects for the given force level as described in Section Appendix F.

Figure A.16: Random Forest Variable Importance Plots (with LEMAS Variables)

Notes: Figure presents variables' average effects on the accuracy of random forest models on classification models for whether a department has a disparity in its use of force against Black subjects for the given force level as described in Section Appendix F.