

Modern Police Tactics, Police-Citizen Interactions, and the Prospects for Reform

Jonathan Mummolo, Princeton University

High-profile incidents of police misconduct have led to widespread calls for law enforcement reform. But prior studies cast doubt on whether police commanders can control officers, and offer few policy remedies because of their focus on potentially immutable officer traits like personality. I advance an alternative, institutional perspective and demonstrate that police officers—sometimes characterized as autonomous—are highly responsive to managerial directives. Using millions of records of police-citizen interactions alongside officer interviews, I evaluate the impact of a change to the protocol for stopping criminal suspects on police performance. An interrupted time series analysis shows the directive produced an immediate increase in the rate of stops producing evidence of the suspected crime. Interviewed officers said the order signaled increased managerial scrutiny, leading them to adopt more conservative tactics. Procedural changes can quickly and dramatically alter officer behavior, suggesting a reform strategy sometimes forestalled by psychological and personality-driven accounts of police reform.

The war on drugs and the adoption of “broken windows” law enforcement tactics (Wilson and Kelling 1982) that aggressively target so-called quality of life crimes have made frequent contact with police a fact of life for millions of Americans, especially Americans of color (Alexander 2010; Gottschalk 2008; Sampson and Loeffler 2010; Travis, Western, and Redburn 2014). As a spate of high-profile episodes of police violence has demonstrated, unlike contact with other bureaucrats, encounters with police present unique psychological and physical risks to the citizen ranging from inconvenience to humiliation, injury, and death. Negative contact with law enforcement has also been shown to depress political participation and erode views of the state (Burch 2013; Lerman and Weaver 2014a, 2014b) and place significant economic burdens on citizens (Howell 2009; Meredith and Morse 2015). Moreover, as the net cast by the criminal justice system has widened in recent decades while crime rates have generally fallen, the correlation between contact with police and criminal guilt has grown “increasingly tenuous” (Lerman and Weaver 2014a, 3). These trends not only harm citizens but inhibit police work, as perceptions of unfair policing diminish support for, and cooperation with, law enforcement (e.g., Tyler

and Wakslak 2004). Calls for reform and oversight of police organizations are now widespread (Martin 2014; Schmidt 2015; Vitale 2014).

But even if social movements aimed at reforming policing garner victories in courts and legislatures, reforms will have to be implemented within police organizations. Decades of research on police misconduct and administration suggests that police managers may find it difficult to control the behavior of their officers. Scholars of organizations and public bureaucracies have long understood management issues in public institutions as principal-agent problems and have debated the degree to which monitoring coupled with the credible threat of sanctions causes workers to comply with managerial directives (Downs 1967; McCubbins, Noll, and Weingast 1987; Miller 2005). While these approaches have proved promising in a number of settings (e.g., Olken 2010), police scholars have long expressed doubts about the ability of rules and supervision to shape officer behavior, citing officer “predispositions” (e.g., Brehm and Gates 1999) and the difficulty of observing police activity as powerful impediments (Davis 1971; Goldstein 1960; Wilson 1968). But prior empirical tests of these claims have been hampered by a

Jonathan Mummolo (jmummolo@princeton.edu) is an assistant professor of politics and public affairs at Princeton University, Princeton, NJ 08544.

This project was part of dissertation work funded by a National Science Foundation dissertation completion grant. Data and supporting materials necessary to reproduce the numerical results in the paper are available in the *JOP* Dataverse (<https://dataverse.harvard.edu/dataverse/jop>). An online appendix with supplementary material is available at <http://dx.doi.org/10.1086/694393>. This research was approved by Stanford University’s Institutional Review Board.

The Journal of Politics, volume 80, number 1. Published online December 6, 2017. <http://dx.doi.org/10.1086/694393>

© 2017 by the Southern Political Science Association. All rights reserved. 0022-3816/2018/8001-00XX\$10.00

000

lack of high-resolution data on officer behavior and research designs that were ill equipped to facilitate valid causal inferences. In addition, many studies in this arena have focused on reducing officer “shirking” and examine productivity-based outcomes such as arrest rates regardless of whether arrests are warranted (Engel 2000), outcomes that are uninformative for investigations centered on the fairness of police-citizen interactions.

This study overcomes these limitations by identifying a rare case where granular, high-frequency records of police officer behavior were recorded before and after an unanticipated procedural reform to a highly controversial tactic, “stop, question, and frisk” (SQF) in New York City. This tactic has been widely criticized as inefficient and overzealously applied and driven by racial profiling (Gelman, Fagan, and Kiss 2007). In March of 2013, the New York Police Department (NYPD) suddenly mandated that officers provide thorough, narrative descriptions to superiors justifying the reasons for stops of criminal suspects. Original interviews show that this directive convinced some officers that commanders would now be further scrutinizing their judgement on the street and were “trying to find a reason to penalize” officers for misconduct surrounding the tactic (officer 1).¹ With this new perceived risk of sanction for questionable stops, many officers began limiting stops to instances where the probability of criminal activity appeared relatively high. Using millions of observations of police-citizen encounters contained in the NYPD’s SQF database, an interrupted time series analysis supports this qualitative account and shows that the rate of stops which produced evidence of the crime suspected by the officer sharply and immediately increased following the new directive. A range of robustness checks reported below indicate that the change is very unlikely to have been caused by reporting bias or data manipulation. Further, contrary to claims that a reduction in SQF activity (due to reforms such as this one) have led to a surge in violent crime in New York City (e.g., Parascandola et al. 2014), I find no discernible change in violent crime following this procedural reform.

While this new directive caused far fewer suspects to be detained by police for crimes they did not commit, this reform was not a panacea. There is no evidence that racial disparities in the rate of stops producing evidence that existed prior to the intervention were eliminated, and while the hit rate improved on average, there is modest evidence

that improvements were most pronounced in neighborhoods with higher shares of white residents. In addition, while the reform did not result in any detectable increase in violent crime or immediate decrease in the number of weapons recovered, it may have resulted in fewer weapons being taken off the street in the months and years that followed. Despite these limitations, the results show that, contrary to prominent claims in the policing literature, officers are highly responsive to rules and supervision, suggesting that institutional changes offer a promising, straightforward avenue for police reform that has been forestalled by a focus in both scholarship and popular discussion on potentially immutable officer traits as the culprits of police behavior and misconduct.

AN INSTITUTIONAL PATH TO POLICE REFORM

Volumes of research in sociology, psychology, and criminology have advanced a “rotten apple” theory of police misconduct (Bonnano 2015) explaining variation in police behavior with individual-level officer traits. For example, scholars have posited that police officers have distinct personalities characterized by “machismo, bravery, authoritarianism, cynicism, and aggression,” as well as bigotry (Balch 1972; Skolnick 1977; Twersky-Glasner 2005, 58), and some argue that police work itself fosters authoritarian personality traits (Laguna et al. 2009; McNamara 1967; Niederhoffer 1967). Police behavior has also been linked to aggressiveness (Hargrave, Hiatt, and Gaffney 1988), conservative ideology (Christie et al. 1995; Fielding and Fielding 1991), and substance abuse (Sellbom et al. 2007). A related strand of research explores the influence of racial bias on police performance and decision making, showing that officers apply lower evidentiary thresholds (Gelman et al. 2007; Glaser 2014; Goel et al. 2016) and a greater propensity to use force (Correll et al. 2007; Eberhardt et al. 2004; Legewie 2016) when dealing with nonwhite suspects.

It seems virtually indisputable that such officer-level traits influence the way that officers do their jobs. But for the reformer, this line of research has, to date, offered few viable policy remedies. Interventions including racial, cultural, and gender-based sensitivity training, as well as calls to diversify police forces, are often proposed (Cioccarelli 1989; Levine et al. 2002; Lockwood and Prohaska 2015, 88; Roberg 1978), but evidence for the effectiveness of these interventions is sparse. Smith (2003) shows that the level of racial diversity in a police force fails to predict the rate of police-caused homicides in that jurisdiction. Christie et al. (1995) show that the effect of training thought to reduce authoritarian and conservative tendencies was eclipsed by the countervailing effects of experience on the job among Australian police.

1. Because of the sensitivity of the subject matter, interviewed officers were granted anonymity. Each officer is referred to by an ID number throughout the text.

When studies do show promising results, they are typically plagued by flawed research designs. In an expansive review of over 985 reports on prejudice reduction efforts in a variety of settings, including police departments, Paluck and Green (2009) noted that “entire genres of prejudice-reduction interventions, including diversity training, educational programs, and sensitivity training in health and law enforcement professions, have never been evaluated with experimental methods” (360). The authors concluded that, “we currently do not know whether a wide range of programs and policies tend to work on average” (357). Thus, while research on the microlevel causes of police performance and misconduct remains vital to the study of criminal justice, and while efforts to reduce prejudice should continue to be developed, this line of research faces limitations when it comes to generating effective and actionable policy solutions.

An alternative approach is offered by a vast, multidisciplinary literature on bureaucracies and organizations. Police officers have long been viewed as street-level bureaucrats (Lipsky 1980), whose preferences may differ systematically from those of their superiors (Brehm and Gates 1999). This presents police managers with a principal-agent problem. Managers have incomplete information as to how their officers spend their shifts and must find ways to ensure their compliance with directives. Scholars of organizations have long debated whether a combination of incentives, monitoring and credible threats of sanctions can shape the behavior of workers even in the face of such difficulties (see Miller [2005] for a review of this expansive literature). In the context of bureaucratic organizations, many scholars have shown that an array of institutional actions, such as rule making, budgeting, and the threat of sanctions, can alter the behavior of bureaucrats (Carpenter 1996; Huber and Shipan 2002; McCubbins et al. 1987; Olken 2010).

Despite such evidence, police scholars have long expressed skepticism when it comes to the efficacy of rules and supervision in police organizations for several reasons. For one, police officers often work out of sight from supervisors and their job often entails dealing with unanticipated events in an array of physical locations, all of which makes verifying noncompliance with directives relatively daunting compared to other bureaucratic settings (Goldstein 1960). Unlike other bureaucrats who handle a repetitive set of office tasks each day, the work of a patrol officer is defined by spontaneity, making it difficult for commanders to craft viable orders and verify that they are followed. In addition, the vagueness of many statutes means that a police administrator’s “ability to control the discretion of his subordinates is in many cases quite limited” (Wilson 1968, 227), especially with regard to “order maintenance” tasks, a category of po-

lice work that includes stopping suspicious individuals for questioning.²

Applying principal-agent theory to police organizations, Brehm and Gates (1999) paint a portrait of a supervisor constrained not only by time and resources but by the dispositions of her workers. “Getting the incentive structure ‘right’ may not be enough,” they write (40). “In prior principal-agent models, one sees compliance from the subordinates if the supervisor’s punishment poses a credible threat. In our model, one sees compliance when subordinate predispositions favor the policy” (44). In an empirical analysis of police brutality, Brehm and Gates (1999) find no evidence that policies or sanctions curbed police violence (168). “We would not go as far . . . to call the coercive power of supervision a ‘fiction,’” the authors conclude, “but the results . . . do suggest that it is an awfully short story” (171).

But while the idiosyncrasies of policing and officers’ preferences surely constrain management, there have been significant changes to the policing environment since the time of several of these foundational studies. The prevalence of smart phone cameras, open data policies (James 2015), and civilian review boards have made police behavior much more visible. Recent decades have also seen an explosion in the rate of police-citizen contacts (Gelman et al. 2007; Goel et al. 2016; Lerman and Weaver 2014a) and high-profile law suits brought by watchdog groups, which have forced improved record keeping inside police agencies. In turn, the perceived threat of sanctions to officers for noncompliance with directives is now arguably more credible than ever, since noncompliance is far more likely to be discovered by management (Fisher and Hermann 2015). These changes, coupled with widespread calls for reform and new high-resolution data sets on police-citizen interactions, necessitate a rigorous empirical reassessment of the responsiveness of police officers to institutional directives.

A SUDDEN PROCEDURAL CHANGE IN NEW YORK CITY

Though it has long been a part of police work, the legal authority for SQF comes from *Terry v. Ohio* (1968), a Supreme Court case that ruled officers who observe articulable facts that are indicative of criminal activity may temporarily detain, question, and potentially search that individual in order to investigate further. As Alexander (2010) notes, since *Terry*, “stops, interrogations, and searches of ordinary peo-

2. As Wilson (1968) states, “to get the patrolman to ‘do the right thing’ when he is making ‘street stops’ . . . the administrator must first be able to tell him what the right thing is. This is seldom possible” (64).

ple driving down the street, walking home from the bus stop, or riding the train, have become commonplace—at least for people of color” (63–64).

In time, “stop, question and frisk,” once considered to be an optional investigative tool, came to be regarded as a measure of an officer’s productivity in the NYPD (Rayman 2013). During much of the 2000s, many officers claimed that failure to report sufficient numbers of stops could result in an array of punishments and career setbacks (Rayman 2013). In this organizational climate, rates of stops by police soared, growing by 603% between 2002 and 2011, reaching nearly 700,000 stops in 2011 (Lerman and Weaver 2014a, 36–37). According to the data analyzed in this study, a summons was not issued and an arrest was not made in nearly 90% of stops made in New York City from 2008 through 2012. In addition, roughly 90% of stopped suspects in that period were nonwhite, though more than 40% of city residents are white.

Stops are recorded in the NYPD on “UF-250” forms, which are filled out by officers. These forms, which at the time under study, consisted of short fields and check boxes (see fig. A2; figs. A1–A5, B1–B5, C1–C3, D1–D5, E1–E5 available online), convey the date, time, and location of each stop, as well as the reason (suspected crime and other circumstances), suspect attributes, and various outcomes such as whether a weapon was found or an arrest was made. Critics of SQF had long alleged that this form was insufficient to establish the legality of a stop. With the trial for a class action law suit concerning the policy, *David Floyd, et al. v. City of New York*, set to commence, plaintiffs in the case filed a memo dated March 4, 2013, in US District Court asking for several reforms, among them the following request: “the UF-250 form should be modified to: (i) include a narrative portion for police officers to justify the basis for stops, frisks, and searches” (Center for Constitutional Rights 2013a, 17). Though NYPD patrol guide documents show that officers had long been required to “enter details” about each stop in their activity logs (notebooks), there was no requirement that these notes be turned in to supervisors along with UF-250 forms after each shift.³ According to media accounts, the plaintiffs did not expect a policy change from the NYPD after filing this brief, since this reform was something they had “been asking for for ten years” without success (Devereaux 2013b).

But on the very next day, March 5, 2013, the NYPD’s then-Chief of Patrol James P. Hall issued a memo to the

3. Periodic audits of officers’ notebooks began “in or around 2008.” Plaintiffs argued these audits showed frequent noncompliance with the order to record the details of stops (transcript from *Floyd*, March 18, 2013 25, 53; Center for Constitutional Rights 2013b).

commanders of all patrol units (see fig. A1), essentially mandating this exact reform. In addition to reinforcing the mandate that officers record notes on the details of stops in their activity logs, the memo contained a new order requiring officers to photocopy and submit these narrative descriptions of the reasons they stopped suspects to supervisors after each shift. As the evidence below will show, this intervention suddenly increased the perceived level of supervision being applied to officers’ decision making on the street.⁴

In court testimony, Hall said the proximity of the memo to the Floyd trial was a coincidence (transcript from *Floyd*, May 16, 2013, 7684) and that the memo was modeled off of a previous memo disseminated in a patrol borough in Queens earlier that year.⁵ But Darius Charney, the plaintiffs’ lead attorney in *Floyd*, called the memo “gamesmanship pure and simple,” since it was released just one day after the plaintiffs’ brief requesting the same reform (Horan 2013). The directive therefore may have been a legal tactic meant to persuade the court that reforms to SQF being sought through litigation could be handled internally (Horan 2013). The fact that the directive may have been a strategic response made to a brief filed just one day earlier is important, since it implies it was not long-planned or anticipated by NYPD officers. This serves to mitigate concerns about anticipatory behavior on the part of officers that could otherwise hamper the unbiased estimation of the directive’s impact.

DATA AND METHODS

The current study is designed to overcome several weaknesses in prior empirical work measuring the ability of rules and supervision to influence officer behavior. Many prior studies focus on reducing “shirking” and examine productivity-based outcomes such as arrest rates (Engel 2000) or the length of police-citizen encounters (Allen 1982) that are of limited use for investigations centered on the fairness with which citizens are treated by police. In addition, studies in this area often don’t leverage randomized or as-if randomized interventions, leaving behind the substantial threat of omitted variable bias.⁶ Finally, a lack of readily available administrative data has led researchers to rely on relatively small convenience samples,

4. According to court testimony from Inspector Juanita Holmes, commanding officer of the department’s 81st Precinct, this new requirement streamlined the supervisory process (transcript from *Floyd*, May 9, 2013, 6546).

5. A request to the NYPD for all SQF-related memos by the Queens North commander in that time period was denied. A request to discuss the results of this analysis made to the NYPD’s public information office on May 11, 2015, was not returned.

6. But see Ariel, Farrar, and Sutherland (2015).

often gathered via participant-observation methods such as “squad car anthropology” (Allen 1982, 92), a technique highly vulnerable to demand effects (Orne 1962).

In contrast, the primary source of quantitative data in the current study is the NYPD’s publicly available SQF database (2008–15), which contains over 3 million records of police-citizen interactions.⁷ This study tests whether a procedural change inside the NYPD increased the citywide rate of stops of criminal suspects that produced evidence of the crime suspected by the officer. In a legal sense, a stop can be justified even if evidence of a crime is not uncovered. But determining whether the officer’s suspicion was correct is important. To the extent that officers make stops because they mistakenly perceive criminal activity or, upon making stops, discover crimes unrelated to their motivating suspicion, SQF can devolve from a potentially useful investigative tool to a frequent, largely arbitrary, and potentially dangerous intrusion into the lives of the policed. In keeping with other recent work, this study therefore uses an outcome-based measure to determine whether the suspicion motivating a stop was accurate (Ayres 2001; Hernandez-Murillo and Knowles 2004; Knowles et al. 2001; Persico and Castleman 2005; Persico and Todd 2006; cited in Engel 2008). Specifically, the dependent variable is an indicator of whether a stop that occurred due to the suspected crime of “criminal possession of a weapon” in fact produced a weapon (Goel et al. 2016), a version of a statistic commonly known as the “hit rate.” This metric conveys both the efficiency and fairness with which the tactic was applied.

This version of the hit rate was chosen for several reasons. First, criminal possession of a weapon (CPW) is the most common suspected crime in the SQF data (it accounts for roughly 26% of stops in pretreatment data) and also corresponds to one of the NYPD’s chief goals for SQF, pulling illegal weapons off the street (Devereaux 2013a).⁸ Achieving a higher weapon recovery rate is therefore theoretically appealing to both police and citizens. Second, unlike other versions of the hit rate, such as arrest rates, this measure explicitly links the suspected crime to tangible evidence of that crime that is difficult to falsify, providing an objective basis for determining whether a police officer’s suspicion was warranted.

To determine whether the memo induced an improvement in the hit rate, this study employs an interrupted time series analysis, a variety of regression discontinuity designs

(RDD) in which the running variable is time (Morgan and Winship 2014; Shadish et al. 2002). The SQF data are ideal for this approach because of the high frequency of measurement and well-defined moment of the intervention—the former alleviates concerns about unobserved confounders which change levels during long intervals between observations, and the latter guards against researcher discretion in coding treated and untreated units. The primary quantity of interest is the immediate change in the probability of recovering a weapon during a stop on March 5, 2013, represented by τ in the following ordinary least squares models:

$$\begin{aligned} \text{weapon}_i &= \alpha + \tau \text{memo}_i + s_j(d_i) + \varepsilon_i, \\ i &= 1, 2, \dots, N; \quad j = 1, 2, \dots, 4. \end{aligned} \quad (1)$$

In equation (1), weapon_i is an indicator of whether each stop, i , resulted in a weapon being discovered, α is an intercept, memo_i is an indicator for an observation falling on the day of the intervention or later, $s_j(d_i)$ are various functions that model time trends on either side of the discontinuity using the running variable, d_i —the distance in days, from the day the memo was issued, which can be positive or negative—and ε_i is an error term. The model is specified to either estimate a simple difference in means (in which case the function $s_j(\cdot)$ simply omits d_i from the model) or to model separate linear, quadratic, or cubic functions on either side of the treatment boundary by interacting various orders of d_i with memo_i .⁹

Given this estimation strategy, the key assumption necessary to attribute any immediate change in the hit rate at the moment of the procedural reform to the reform itself is that no other factor which affects the hit rate also systematically changed at the same point in time.¹⁰ If the data had been aggregated by month or year, as is often the case with administrative records, this would be a strong assumption, as we would undoubtedly be conflating myriad events in the time series with the introduction of the treatment. But given the granularity of the SQF data, we are able to isolate the change in the hit rate on the specific day of the reform and can therefore make the much more plausible assumption that leading candidate omitted variables such as criminal activity in the city and department personnel are not also changing suddenly on March 5, 2013.

In order to estimate these models, two broad strategies are applied. The first uses data on all weapon stops from 2008

7. The authors of Goel et al. (2016) generously shared their merged SQF data file covering the period through 2013. Data on 2014 and 2015 were appended. See the appendix for details on data cleaning and merging.

8. Most weapons recovered via SQF are knives. Among weapon stops, less than 12% of stops that produced a weapon in the pretreatment period yielded firearms.

9. For example, the linear model is specified as $\text{weapon}_i = \alpha + \tau \text{memo}_i + \beta_1 d_i + \beta_2 \text{memo}_i * d_i + \varepsilon_i$. See app. A for more details on model specifications.

10. More formally, we must assume continuity in the potential outcome functions at the treatment boundary (de la Cuesta and Imai 2016).

through 2015, an approach that enhances the precision of estimates due to the large sample size but runs the risk of omitted variable bias since it allows other events in the time series to influence estimates of the treatment effect. The second approach aims to minimize such bias by subsetting to a narrow temporal sliver of observations before and after the memo was released, thereby eliminating the influence of events far from the intervention date. The trade-off of this second approach is that the sample size is greatly decreased, meaning that estimates will be less precise. To assure results are not being driven by particular modeling choices, models using narrow temporal windows are estimated using several model specifications and bandwidths (Eggers et al. 2015; Gelman and Imbens 2014; Hall 2015; Imbens and Lemieux 2008).¹¹ Additionally, when all weapon stops are used and the risk of bias is more pronounced, models control for potential time-varying confounders using year, month, and day-of-week indicators, and the prior day’s hit rate.¹²

Quantitative analyses were supplemented with several qualitative sources. The first was a set of court transcripts from the aforementioned class action trial, *David Floyd et al. v. The City of New York*, in which the intervention was discussed in detail. These transcripts are particularly valuable since they were produced just weeks after the intervention and provide testimony given under the penalty of perjury. Phone interviews were also conducted in early 2015 with six NYPD officers who worked for the department during the time of the intervention. This is obviously not a large or representative sample of NYPD personnel—the sample size was limited by the difficulty of locating individuals who would speak candidly about a controversial policy—but does include both uniformed and plain-clothes officers, as well as officers from various types of units (e.g., patrol, street narcotics). These conversations proved illuminating, provided a working knowledge of the process of making and recording stops from an officer’s perspective, and shed light on potential causal mechanisms.

RESULTS

Of the nearly 3.2 million stops recorded from 2008 through 2015, close to 830,000 listed CPW as the suspected crime. For the aforementioned reasons, the core analysis below is performed on these roughly 830,000 observations. Of these

11. Similar results using the *rdrobust* R function developed by Calonico, Cattaneo, and Titiunik (2014), which applies an “optimum” bandwidth, appear in the appendix.

12. Throughout this study, 95% confidence intervals were computed using the maximum of conventional and robust HAC standard errors (Andrews 1991), unless otherwise stated. See the appendix for results with standard errors clustered by precinct.

stops, around 3.5% produced a weapon in the pretreatment period on average, and this rate was remarkably stable for several years leading up to the intervention. However, as figure 1 shows, the hit rate appears to have increased discontinuously on the day of the intervention, and ascended sharply in the months that followed. This visualization provides striking prima facie evidence that the new directive causally affected officer behavior. As Shadish et al. (2002) note, when effects in an interrupted time series are, “immediate and dramatic . . . most threats to internal validity are usually implausible” (176).

In addition, as figure 2 makes plain, the increase in the hit rate at the treatment threshold was caused by an immediate drop in the number of weapon stops being performed (the hit rate’s denominator), not by an increase in the number of stops producing a weapon (the hit rate’s numerator), a fact that will inform the discussion of the likely causal mechanism below. Finally, it is also worth noting that considerable temporal variation in the total number of stops being made in the pretreatment period did not correspond to meaningful changes in the hit rate. For example, though the total number of weapon stops sharply declined in 2012, just before the in-

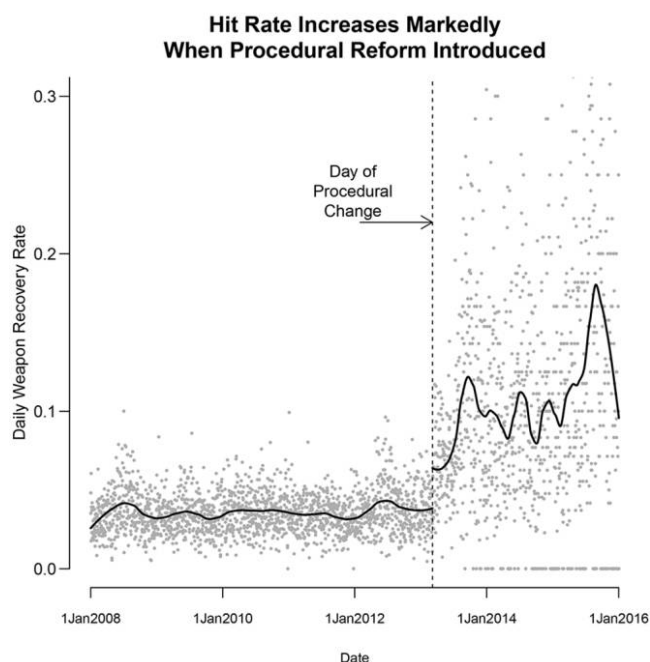


Figure 1. Daily share of stops from 2008 through 2015 in which a weapon was found on a suspect among stops where “criminal possession of a weapon” was listed as the suspected crime. The solid curve is the predicted level of this weapon recovery rate generated by locally weighted (LOESS) regression of daily weapon recovery rates on sequential day numbers, with no adjustment for covariates. Subsequent models use the stop as the unit of analysis. Stop-level data are aggregated as day-level means here to facilitate visualization. For clarity, the y-axis is trimmed and displays the bottom 99% of the data.

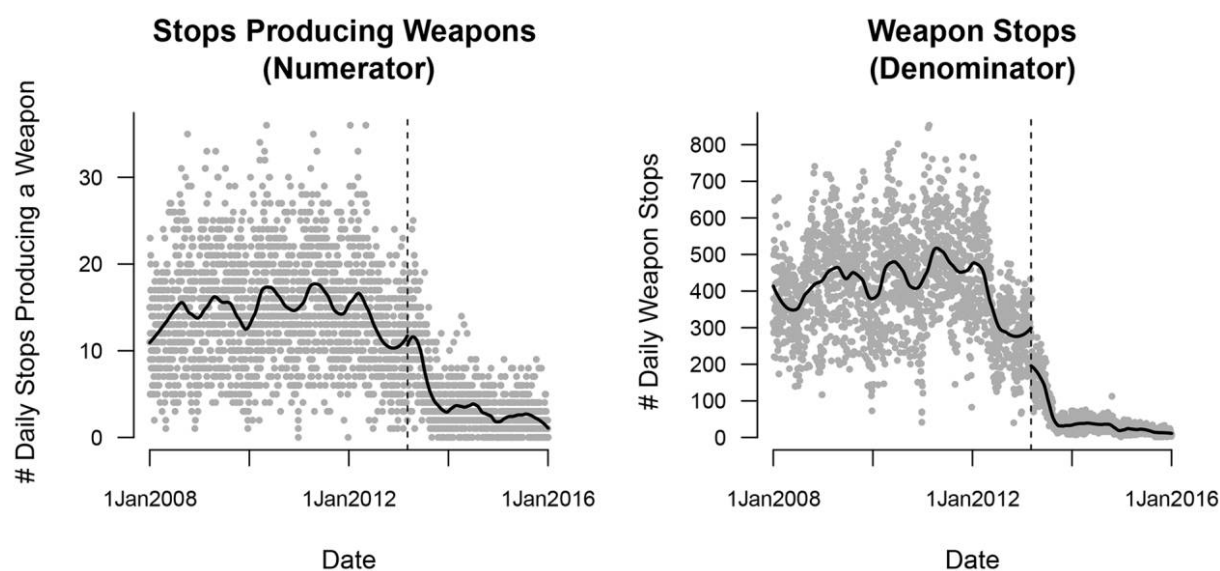


Figure 2. *Left*, Daily number of weapon stops producing a weapon between 2013 and 2015 (the hit rate's numerator). *Right*, Daily number of weapon stops conducted during the same period (the hit rate's denominator). A discontinuous drop occurs in the hit rate's denominator the day of the intervention.

intervention—possibly due to mounting controversy surrounding the tactic which caused supervisors to relax demands for stops (Goldstein and Ruderman 2012)—the hit rate remained more or less stable. This is important, since it demonstrates that factors which led officers to simply make fewer stops were insufficient to improve the rate of stops producing evidence of the suspected crime. This suggests that simply removing the alleged quota system reportedly in operation during the pre-treatment period (i.e., allowing officers to make fewer stops), without imposing the reporting requirements contained in this directive, would have been insufficient to increase the hit rate on its own.

The first set of formal tests of whether this apparent discontinuity is discernible from zero are displayed in table 1. These OLS estimates, fit to the entire corpus of weapon stops from 2008 through 2015, feature four functional forms with various levels of flexibility. The first two models estimate the mean difference in the probability of discovering a weapon before and after the intervention, while the remaining six models fit separate linear, quadratic, and cubic functions to the data on either side of the intervention date (i.e., three other forms of $s_j(\cdot)$ in eq. [1] above). As the table shows, across a variety of specifications, the probability of recovering a weapon during a given stop is estimated to increase by anywhere between 1 and 5 percentage points, all statistically significant—and substantively large—changes given the pretreatment baseline of 3.5%.

The next set of tests were conducted on narrow bandwidths of data in order to reduce bias while avoiding model-dependent results that hinge on the bandwidth chosen by the

researcher. Figure 3 displays the estimated treatment effects using six different functional forms.¹³ Similar to the results using all weapon stops, the estimated treatment effects using only data close to the date of the intervention are concentrated in the range between roughly 1 and 5 percentage points. Because far fewer observations are being used for estimation, the confidence intervals are larger, especially for the highly flexible estimators. Despite this, the estimates are still statistically significant in many cases, and across all tests there is not a single negative point estimate. In sum, there is robust evidence for a large and immediate improvement in the weapon recovery rate the day of this procedural change.

THE MECHANISM BEHIND IMPROVED PERFORMANCE

With well over 30,000 officers, the NYPD is the nation's largest municipal police department, and the volume of memos circulating its halls is considerable. What about this particular memo caused such an abrupt change in officer behavior? A plausible explanation was supported by officer interviews: the memo increased the perceived probability of being scrutinized and sanctioned for making a wrongful stop, leading to a more conservative use of the tactic.¹⁴ Whereas before the intervention officers occasionally faced

13. In very narrow bandwidths, not all parameters could be estimated in some model specifications because the covariate matrix was not of full rank. Estimates are omitted from fig. 3 in these cases.

14. While interviewed officers were not unanimous on this point, it was a recurring theme in several interviews.

Table 1. OLS Estimates of Discontinuity, All Weapon Stops 2008–15

	Difference in Means	Difference in Means ^a	Linear	Linear ^a	Quadratic	Quadratic ^a	Cubic	Cubic ^a
$\hat{\tau}$.051*	.031*	.030*	.022*	.029*	.020*	.013*	.010*
	(.002)	(.003)	(.002)	(.003)	(.003)	(.003)	(.003)	(.004)
N	826,573	826,260	826,573	826,260	826,573	826,260	826,573	826,260

Note. Maximum of homoskedastic and HAC standard errors in parentheses.

^a Includes controls for year, month, day of week, and prior day’s hit rate.

* $p < .05$, two-tailed.

scrutiny over SQF by those outside the department via suspect complaints and law suits, the memo signaled to officers that heightened scrutiny would now be coming from supervisors. “They’re really watching us now,” one officer recalled thinking when the memo was released (officer 2).

Another officer added that before the memo, supervisors “would only look at [memo book entries] if someone made an allegation . . . or you had to go to court . . . Now . . . it’s basically like they’re looking at it . . . without any sort of allegation being made . . . They’re trying to find a reason to

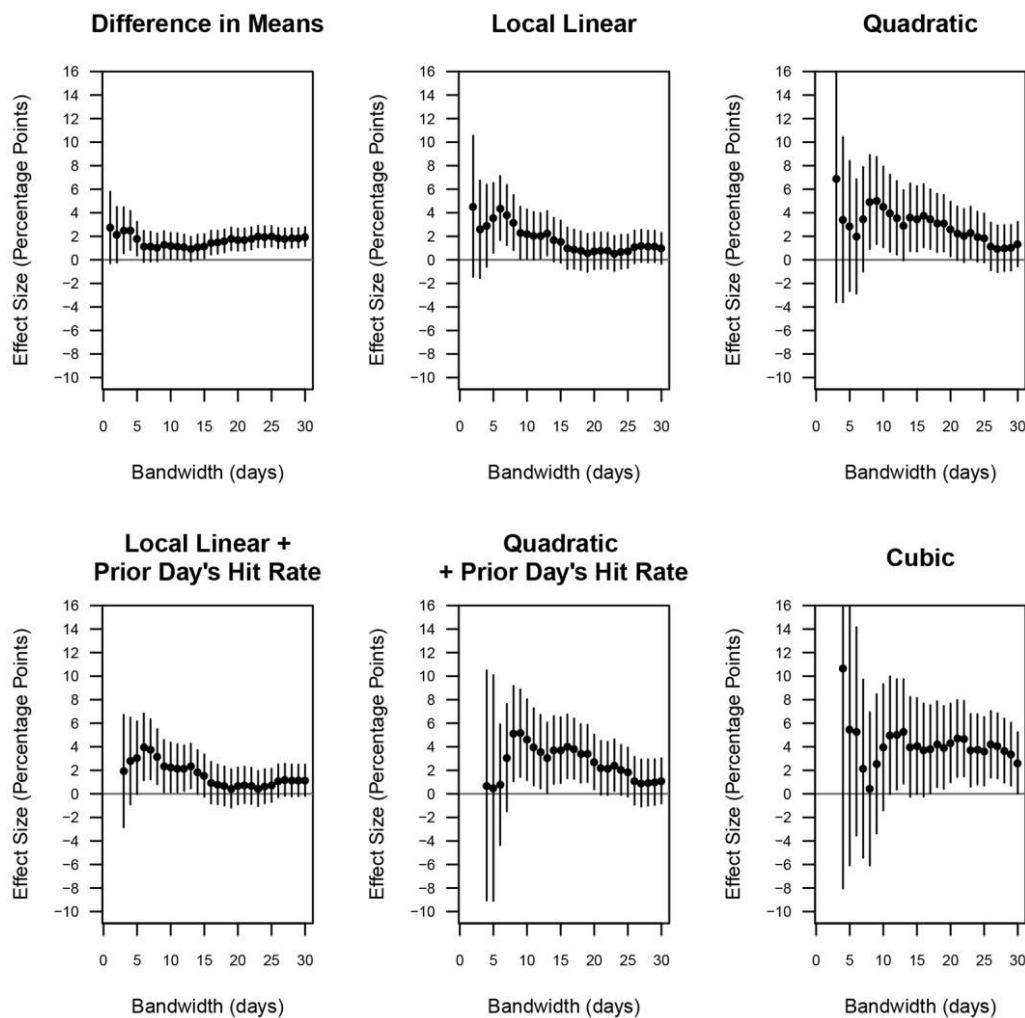


Figure 3. Estimates of the change in probability of recovering a weapon the day of the intervention using various model specifications and bandwidths. Vertical lines denote 95% confidence intervals for each estimate.

Table 2. OLS Estimates of Change in Daily Stops Producing a Weapon (“Hits”) and Daily Weapon Stops (“Stops”) at Treatment Threshold Using 100-Day Bandwidth

	Difference in Means	Difference in Means ^a	Linear	Linear ^a	Quadratic	Quadratic ^a	Cubic	Cubic ^a
Hits:								
Δ	.420 (.800)	1.235 (2.216)	-.260 (1.416)	1.881 (2.323)	-1.081 (2.144)	.165 (2.520)	.653 (2.869)	1.400 (2.710)
Stops:								
Δ	-108.33* (13.311)	-60.989* (25.281)	-120.564* (26.917)	-55.582* (26.177)	-109.008* (43.595)	-50.481 (29.615)	-39.737 (41.204)	-30.83 (32.656)

Note. Maximum of homoskedastic and HAC standard errors in parentheses. $N = 200$.

^a Includes controls for year, month, day of week, and prior day’s hit rate.

* $p < .05$, two-tailed.

penalize us” (officer 1). Supervisors “obviously look at these things with a fine-tooth comb,” said another officer. “We need to protect ourselves” (officer 3).

According to interviews, this perception of increased risk led some officers to aggressively forego making stops unless they observed something highly incriminating. “It’s forcing people to not get involved in things that otherwise, a few years ago, they would have,” said one officer (officer 3). But while the incentives to make low-probability stops were perceived as declining, incentives to make stops in which the officer’s suspicion of criminal activity was likely to be validated by the outcome remained, especially for weapon-related stops. According to interviews, one of the most prestigious achievements in the NYPD is to pull an illegal firearm off the street. “I’ve often heard bosses and cops judge a unit based on how many gun collars they get,” said one officer (officer 6).

The immediate decrease in the number of weapon stops at the treatment boundary portrayed in figure 1—and the apparent stability of the number of stops producing a weapon at that point in time—is consistent with the intervention causing officers to avoid stops with a low probability of a hit. To quantify these changes, I summed both quantities at the day level and estimated the six model specifications used in the global discontinuity models, and the results are displayed in table 2.¹⁵ While the number of weapon stops dropped by as many as 120 the day the memo was released, the number of stops producing a weapon remained stable on that day.

15. Because aggregating the data by day drastically reduces the sample size, I used a 100-day bandwidth for these models. In models with controls, the lagged hit rate is replaced with the lagged number of hits or stops, respectively.

It is worth noting that in more recent years, the number of weapons recovered has fallen to low levels, as figure 1 shows. We cannot credibly attribute this decline to the reform in question, since intervening events such as the NYPD’s loss in the *Floyd* case, the departure of Mayor Michael Bloomberg and the policies of a new police commissioner could all be responsible for that subsequent decline. Only changes at the treatment threshold can be attributed to the intervention with reasonable confidence. However, we also cannot rule out the possibility that the reduction in recovered weapons that occurred in the months and years following the intervention was due to a lagged treatment effect. What we can say is that in the short term, where we have the most leverage for a valid causal inference, this intervention appears to have spared many individuals from being needlessly investigated by police while doing little to impede the recovery of weapons.

Of course, the recovery of weapons is just one measure of the public safety impact of this intervention. Some critics have argued that the decline of SQF in New York has led to increases in violent crime (Parascandola et al. 2014, 2015). Prior research shows little evidence for this claim. Using precinct-year panel data, Rosenfeld and Fornago (2012) find no robust relationship between stop activity and burglaries and robberies within precincts. An internal NYPD report came to a similar conclusion regarding shootings (Parascandola 2015).¹⁶

16. MacDonald, Fagan, and Geller (2016) conclude that “saturating high crime blocks with police helped reduce crime in New York City, but that the bulk of the investigative stops did not play an important role in the crime reductions” (1).

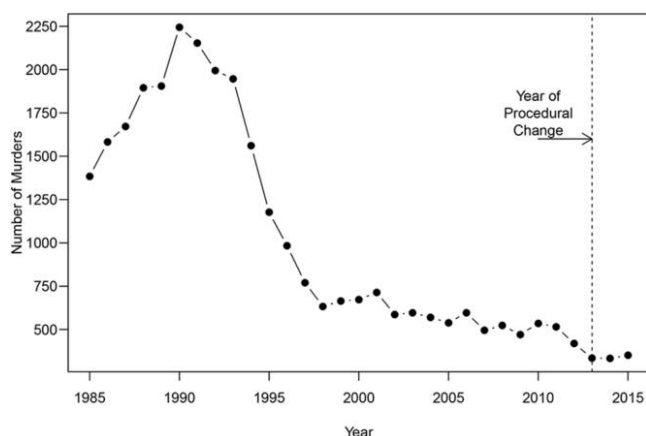


Figure 4. Annual number of homicides in New York City since 1985. Source: FBI Uniform Crime Reports and NYPD.

Homicides are a useful measure of crime to examine when assessing how crime changed post-treatment, since the severity of the crime ensures that the rates at which they are recorded by the city should not depend on the fact that fewer stops were being made via SQF. As figure 4 shows, homicide levels have remained near a decades-long low since the year of the intervention. For a more fine-grained test of whether this procedural reform affected violent crime in the city, I examine weekly homicides around the time of the intervention.¹⁷ As figure 5 shows, there is no evidence of a discontinuous increase in homicides at the moment of the intervention. While we cannot rule out whether homicides would have been even lower after the intervention had the rate of stops been maintained, in the weeks that followed, the overall level of homicides remained at or below the typical range in data going back to 2010, especially once seasonality is accounted for. In short, we see no evidence of a surge in violent crime.¹⁸

HETEROGENEOUS TREATMENT EFFECTS

Exploring the heterogeneity of these effects is especially important in light of allegations of racially biased policing practices, and given that crime rates vary markedly within New York City. I therefore estimated the differences in treatment effects between census block groups with high and low shares of white residents, precincts with high and low homicide rates, and between stops made of white and nonwhite suspects.¹⁹ Using the full corpus of weapon stops,

17. See app. A for information on the data quality of weekly crime statistics.

18. See table B4 in the appendix for the results of formal tests for a discontinuity.

19. Census block groups and precincts at or above the median are coded as “high.”

there was some evidence that treatment effects were larger in low-crime precincts, in block groups with higher shares of white residents and among white suspects. But estimates in local bandwidths, where the potential for bias is reduced, were extremely imprecise, making it difficult to draw firm conclusions (see app. C; apps. A–E available online). Still, these results highlight that this intervention was not a panacea: it likely did not improve the hit rate to the same extent in various locations and for various groups of suspects, and there is no evidence it closed the historic disparities in the hit rate across racial groups that previous scholars have cited as evidence of racially biased policing.

REPORTING BIAS

Reporting bias is always a concern whenever records are generated by those who stand to benefit from their content (McCubbins et al. 1987). However, not all forms of reporting bias are problematic for this study. For example, if the rate of some type of data manipulation remains constant at the treatment boundary, it would not impose bias. However, if immediately after the intervention, in the absence of finding a weapon, officers began to apply a different suspected crime category (other than CPW) to stops that would previously have been labeled weapon stops, the weapon recovery rate could be artificially inflated. Fortunately, this type of behavior should be observable.

First, if reclassification of this sort were occurring, we might expect to see an increase in the frequency of stops labeled under some other crime category after the intervention. But as the top left panel of figure 6 shows, the daily frequency of stops across all suspected crime categories declined with the intervention. Second, reclassification of failed stops should lead to a decrease in the hit rate among *nonweapon* stops, since such behavior would flood the denominator of the hit rate among nonweapon stops with “misses.” The middle four panels of figure 6 show that nearly all point estimates of the discontinuity at the treatment threshold among nonweapon stops are at or near zero, indicating no change in the weapon recovery rate among these stops.²⁰ Third, while it may be plausible that officers began to systematically recode the suspected crime field of their forms after the intervention, it is highly unlikely that officers would be able to correctly adjust the levels of the correlates of the suspected crime category (e.g., suspect race, suspect age, location type, whether a suspicious object was observed), so as to preserve all covariances in the data and mask their behavior. If this multivariate reclassification is not occurring but crime category reclassification is present, then

20. The hit rate among nonweapon stops in the pretreatment period was 0.4%.

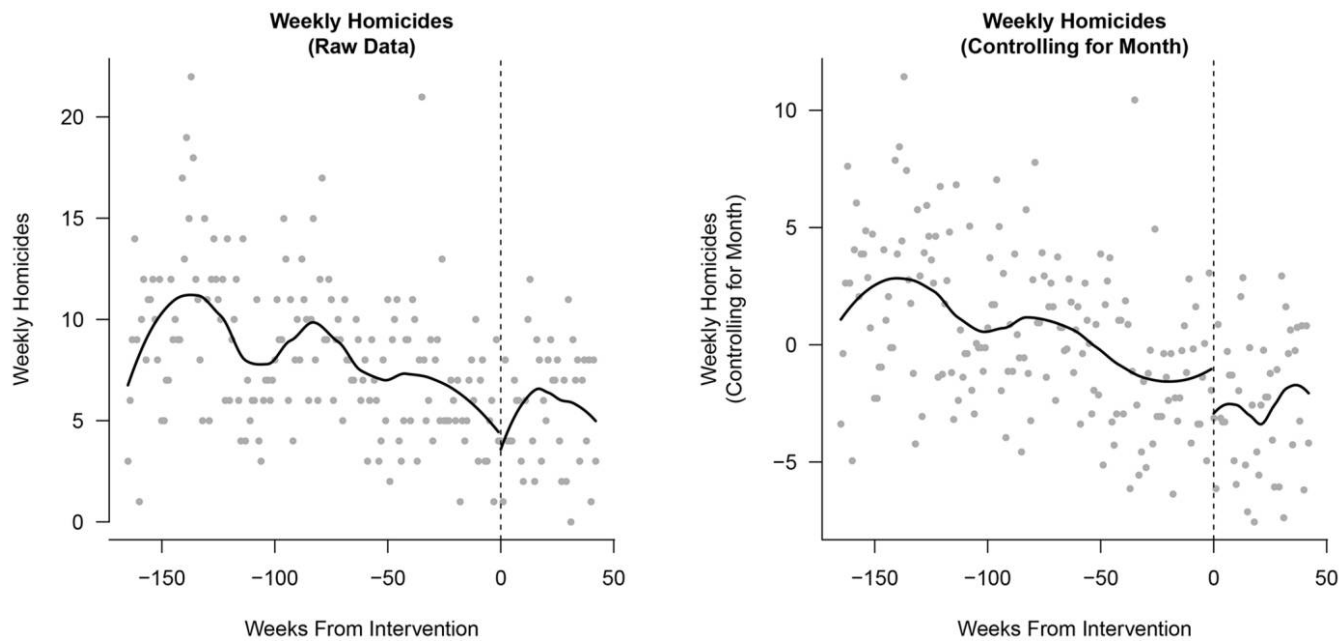


Figure 5. LOESS predictions of weekly homicide counts in New York City between 2011 and 2013. Vertical dotted line denotes the week of the intervention. *Left*, Raw counts; *right*, counts after residualizing the homicide data with respect to month indicators.

nonweapon stops just after the intervention should look more like weapon stops than the nonweapon stops just before the intervention, in terms of their covariates. This change should be reflected in the distributions of predicted probabilities of being labeled a weapon stop among pre- and -post treatment observations, as generated by a logit model predicting having CPW as the suspected crime.²¹ The top right panel of figure 6 shows that, based on the levels of their covariates, nonweapon stops just before and after the intervention had nearly the same predicted probabilities of being labeled a weapon stop. Taken together, the results strongly refute the reclassification hypothesis.

Another form of reporting bias would occur if officers started to hide failed stops from supervisors after the intervention altogether.²² Though some stops surely go unreported, there are reasons to suspect the rate of this behavior did not increase with the intervention. If the desire to avoid discipline was motivating officers, failing to report stops—that is, lying to supervisors—would potentially be a larger risk than reporting a stop that did not produce evidence of a crime. This is especially the case since, as one interviewed officer noted, “there’s basically a camera on every block,” in New York (officer 3). It also requires more labor and physical risk to make and hide stops than to not make stops in the first place.

21. See app. D for details on how these probabilities were generated.

22. It would also be problematic if officers had been reporting stops which never occurred prior to the intervention and curtailed this activity on March 5, 2013. See fig. D4 for evidence against this hypothesis.

In addition, interviewed officers agreed that plain clothes officers were more likely to fail to report stops than uniformed officers who have their identities on display to stopped suspects. When officers’ identities are on display, interviewed officers said, suspects who feel they were mistreated are able to make much more credible complaints to the city. If, in the course of investigating that complaint, it is discovered that no stop was ever reported by the officer, it is very likely that the officer will be found to be at fault and face disciplinary action, interviewed officers said. This suggests an additional robustness check: estimating the hit rate among stops made by officers in uniform (roughly 70% of the data), where the chances of this sort of data censoring are very low. When this is done, estimated treatment effects, displayed in the bottom four panels of figure 6, look nearly identical to those in the full sample. While data censoring cannot be completely ruled out, there is little indication that it is responsible for the observed increase in the hit rate.

DISCUSSION AND CONCLUSION

As Lerman and Weaver (2014a) note, being “stopped by foot patrols as they make their way to jobs and homes” is one of the most common ways in which law-abiding citizens come into contact with law enforcement agents (30). The expansion of this practice has led to a host of negative social and political consequences, including depressed political participation, an erosion of trust in law enforcement and a rapid growth in the size of the carceral state (Alexander 2010; Burch 2013; Lerman and Weaver 2014b; Tyler and Fagan

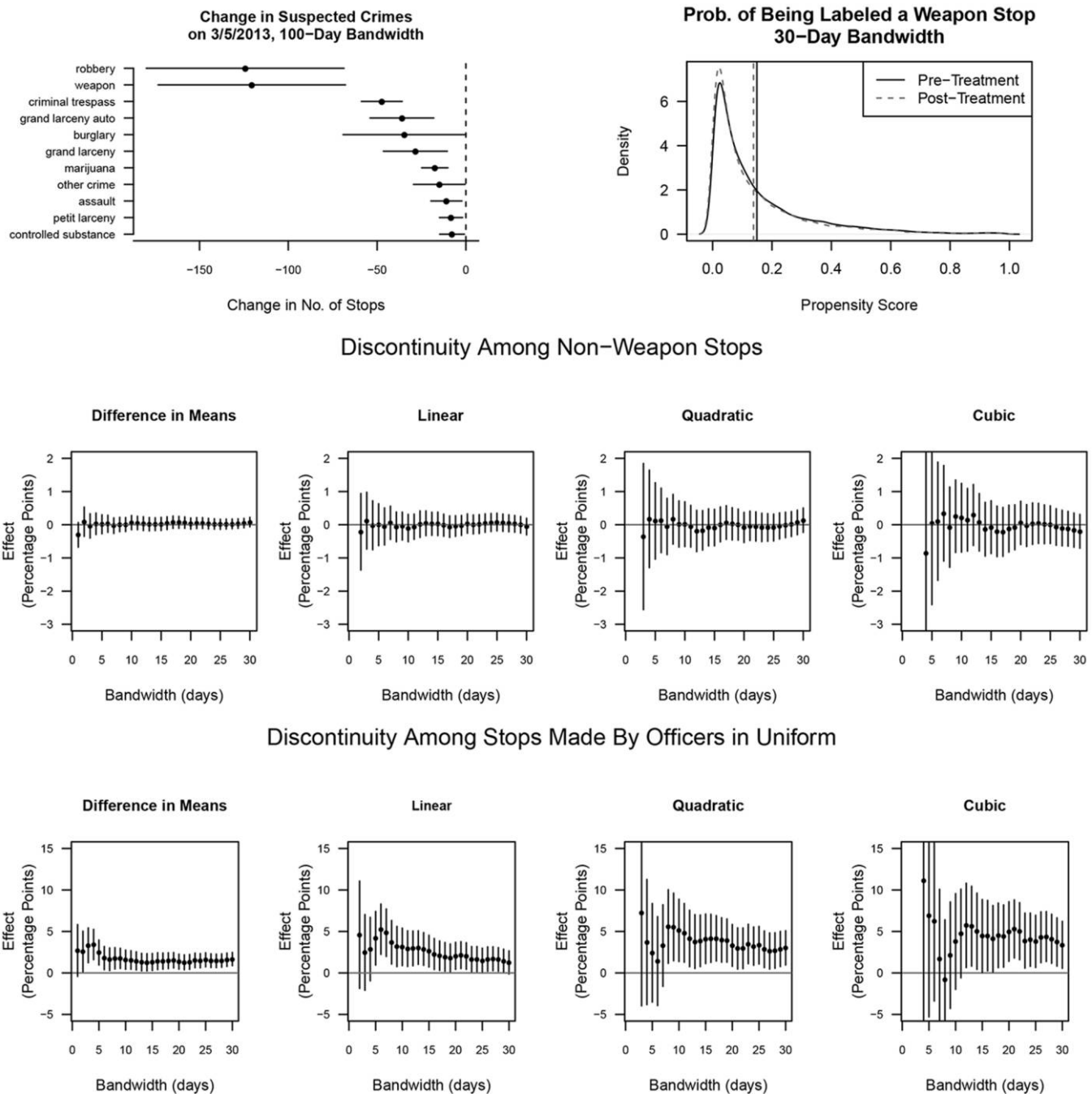


Figure 6. *Top left*, Reduction in the number of daily stops in each suspected crime category at the treatment threshold as estimated by a linear model. *Top right*, Pre- and post-treatment predicted probabilities of being labeled a weapon stop among nonweapon stops, generated by a logit model (vertical lines are means). *Middle*, Estimated discontinuity in the hit rate among nonweapon stops using various model specifications and bandwidths. *Bottom*, Estimated discontinuity in the hit rate among weapon stops made by officers in uniform.

2008). The results above indicate that a simple procedural change to the protocol for reporting the reasons for stops reduced the rate of unnecessary police-citizen interactions. These results are consistent across several estimation strategies and are buttressed by qualitative evidence from a variety of sources, including some of the very NYPD officers who experienced this shift in policy. These results have

significant implications for police reform. Despite persistent claims that police officers are largely autonomous actors who can shirk their duty and defy directives with impunity, we observe instead an immediate change in officer behavior in response to a relatively modest procedural change. This suggests that an array of institutional changes could produce desirable outcomes in terms of police-citizen interactions,

despite the obvious force of the officer traits and preferences which have been the focus of so much prior work and debate surrounding police behavior and misconduct.

There are, of course, some necessary caveats. Though analyzing the immediate discontinuity in the hit rate at the moment of the intervention provides valuable causal leverage, it also confines inferences about this intervention's effectiveness to the short term. The high hit rate observed post-intervention persists and appears to grow through the end of 2015. But we cannot attribute this persistence to the new directive with much confidence, as intervening events could be responsible. This study also examines data from a single city, and the efficacy of similar reforms should be tested and validated in other settings. Future work that selectively implements similar interventions experimentally across multiple departments could test the robustness and persistence of these effects.

The intervention was also followed by a sharp reduction in the number of stops producing a weapon. While we cannot attribute this change to the reform with confidence—since there was no immediate change in this outcome at the treatment boundary, and intervening events could have easily been responsible for future changes—we also cannot rule out the possibility that this reduction was due to a lagged treatment effect. If the treatment did cause this decline, that would represent an important public welfare trade-off. However, it is also worth noting that the primary purpose of removing weapons from the street according to proponents of SQF is to reduce violent crime. As the results show, the intervention did not lead to any detectable increase in homicides or robberies, a result that is consistent with earlier work finding no robust evidence that increases in SQF activity reduced crime rates in New York (Rosenfeld and Fornago 2012).

Despite the stark impact of this reform, the difficulty of improving the quality of police-citizen interactions should also not be understated. Officers still enjoy immense power and discretion as well as substantial barriers to prosecution in the event of wrongdoing (Alexander 2010; Lerman and Weaver 2014a). The effect observed here is limited to a single aspect of police work, and it is possible that performance of other tasks that do not generate reports—or ones performed in environments where the press and public are less able to scrutinize police behavior—would be much more difficult to improve. And even if similar interventions lead to widespread improvements in policing nationwide (a best-case scenario), it may still take years, if not decades, to rebuild the atrophied levels of trust between residents of overpoliced communities and law enforcement personnel.

But as solutions to the problems facing law enforcement continue to be sought, these findings should underscore for reformers the strong influence of institutional factors on

police behavior. The trope of the “rogue cop” in discussions surrounding police misconduct has led to an individuation of social justice problems that, to a large extent, have institutional support. To be clear, this article does not dispute that individual-level factors such as racial bias and personality affect police-citizen interactions but rather that such results, at present, suggest few policy-based remedies. Even if some prejudice reduction strategies are effective, police organizations have often failed to demonstrate this by scientifically evaluating them during implementation. Indeed, the failure to adequately assess the merit of these initiatives may indicate a willful ignorance and illustrate the resistance of institutions to more sweeping structural remedies (Paluck and Green 2009, 343–44). Announcing prejudice reduction initiatives while failing to properly evaluate them may allow political leaders to appear concerned about injustice while distracting attention from the fact that the institutions they control play a substantial role in shaping police behavior.

ACKNOWLEDGMENTS

I am grateful to Justin Grimmer, Amy Lerman, Terry Moe, Gary Segura, Sharad Goel, Jens Hainmueller, Andy Hall, Neil Malhotra, Dan Butler, Lauren Davenport, Jeffrey Fagan, Gary Cox, Dan Hopkins, Erik Peterson, Dorothy Kronick, Sean Westwood, Yiqing Xu, Bobby Gulotty, Ariel White, and Lauren Wright for invaluable advice and support throughout this project.

REFERENCES

- Alexander, Michelle. 2010. *The New Jim Crow: Mass Incarceration in the Age of Colorblindness*. New York: New Press.
- Allen, David N. 1982. “Police Supervision on the Street: An Analysis of Supervisor/Officer Interaction during the Shift.” *Journal of Criminal Justice* 10 (2): 91–109.
- Andrews, Donald W. K. 1991. “Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation.” *Econometrica* 59 (3): 817–58.
- Ariel, Barak, William A. Farrar, and Alex Sutherland. 2015. “The Effect of Police Body-Worn Cameras on Use of Force and Citizens’ Complaints against the Police: A Randomized Controlled Trial.” *Journal of Quantitative Criminology* 3 (31): 509–35.
- Ayres, Ian. 2001. *Pervasive Prejudice? Unconventional Evidence of Racial and Gender Discrimination*. Chicago: University of Chicago Press.
- Balch, Robert. 1972. “Police Personality: Fact or Fiction?” *Journal of Criminal Law and Criminology* 63 (1): 106–19.
- Bonnano, Emma Rose. 2015. “An Evidential Review of Police Misconduct.” 2015 Undergraduate Awards 9. http://ir.lib.uwo.ca/cgi/viewcontent.cgi?article=1008&context=ungradawards_2015.
- Brehm, John, and Scott Gates. 1999. *Working, Shirking, and Sabotage: Bureaucratic Response to a Democratic Public*. Ann Arbor: University of Michigan Press.
- Burch, Traci. 2013. *Trading Democracy for Justice: Criminal Convictions and the Decline of Neighborhood Political Participation*. Chicago: University of Chicago Press.

- Calonico, Sebastian, Matias D. Cattaneo, and Rocio Titiunik. 2014. "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs." *Econometrica* 82 (6): 2295–326.
- Carpenter, Daniel P. 1996. "Adaptive Signal Processing, Hierarchy, and Budgetary Control in Federal Regulation." *American Political Science Review* 90 (2): 283–302.
- Center for Constitutional Rights. 2013a. "Floyd, David, et al. v. The City of New York: Memorandum of Law in Support of Plaintiff's Request for Injunctive Relief." <https://ccrjustice.org/home/what-we-do/our-cases/floyd-et-al-v-city-new-york-et-al> (accessed June 23, 2015).
- Center for Constitutional Rights. 2013b. "Floyd v. New York City Trial Updates." <http://ccrjustice.org/floyd-v-new-york-city-trial-updates> (accessed November 28, 2015).
- Christie, Gayre, Simon Petrie, and Perri Timmins. 1995. "The Effect of Police Education, Training, and Socialisation on Conservative Attitudes." *Australian and New Zealand Journal of Criminology* 29 (3): 299–314.
- Cioccarelli, P. 1989. "Police Education Training." *National Police Research Unit Review* 5:33–45.
- Correll, Joshua, Bernadette Park, Charles M. Judd, and Bernd Wittenbrung. 2007. "The Influence of Stereotypes on Decisions to Shoot." *European Journal of Social Psychology* 37 (1): 1102–17.
- Davis, Kenneth C. 1971. *Discretionary Justice: A Preliminary Inquiry*. Urbana: University of Illinois Press.
- de la Cuesta, Brandon, and Kosuke Imai. 2016. "Misunderstandings about the Regression Discontinuity Design in the Study of Close Elections." *Annual Review of Political Science* 19:375–96.
- Devereaux, Ryan. 2013a. "New York's Stop-and-Frisk Trial Comes to a Close with Landmark Ruling." *The Guardian*, August 12. <http://www.theguardian.com/world/2013/aug/12/stop-and-frisk-landmark-ruling>.
- Devereaux, Ryan. 2013b. "NYPD Stop-and-Frisk Memo Revealed in Civil Rights Court Battle." *The Guardian*, March 27. <https://www.theguardian.com/world/2013/mar/27/nypd-stop-and-frisk-memo>.
- Downs, Anthony. 1967. *Inside Bureaucracy*. Boston: Little, Brown.
- Eberhardt, Jennifer L., Phillip Atiba Goff, Valerie J. Purdie, and Paul G. Davies. 2004. "Seeing Black: Race, Crime, and Visual Processing." *Journal of Personality and Social Psychology* 87 (6): 876–93.
- Eggers, Andrew C., Anthony Fowler, Jens Hainmueller, Andrew B. Hall, and James M. Snyder. 2015. "On the Validity of the Regression Discontinuity Design for Estimating Electoral Effects: New Evidence from over 40,000 Close Races." *American Journal of Political Science* 59 (1): 259–74.
- Engel, Robin S. 2000. "The Effects of Supervisory Styles on Patrol Officer Behavior." *Police Quarterly* 3 (3): 262–93.
- Engel, Robin S. 2008. "A Critique of the 'Outcome Test' in Racial Profiling Research." *Justice Quarterly* 25 (1): 1–36.
- Fielding, Nigel G., and Jane Fielding. 1991. "Police Attitudes to Crime and Punishment: Certainties and Dilemmas." *British Journal of Criminology* 31 (1): 39–53.
- Fisher, Marc, and Peter Hermann. 2015. "Did the McKinney, Texas, Police Officer Know He Was Being Recorded?" *Washington Post*, June 8.
- Gelman, Andrew, and Guido Imbens. 2014. "Why High-Order Polynomials Should Not Be Used in Regression Discontinuity Designs." Working paper no. 20405, National Bureau of Economic Research, Cambridge, MA.
- Gelman, Andrew, Alex Kiss, and Jeffrey Fagan. 2007. "An Analysis of the New York City Police Department's 'Stop-and-Frisk' Policy in the Context of Claims of Racial Bias." *Journal of the American Statistical Association* 102 (479): 813–23.
- Glaser, Jack. 2014. *Suspect Race: Causes and Consequences of Racial Profiling*. Oxford: Oxford University Press.
- Goel, Sharad, Justin M. Rao, and Ravi Shroff. 2016. "Precinct or Prejudice? Understanding Racial Disparities in New York City's Stop-and-Frisk Policy." *Annals of Applied Statistics* 10 (1): 365–94.
- Goldstein, Joseph. 1960. "Police Discretion Not to Invoke the Criminal Process: Low-Visibility Decisions in the Administration of Justice." *Yale Law Journal* 69 (1): 543–88.
- Goldstein, Joseph, and Wendy Ruderman. 2012. "Street Stops in New York Fall as Unease Over Tactic Grows." *New York Times*, August 3.
- Gottschalk, Marie. 2008. "Hiding in Plain Sight." *Annual Review of Political Science* 11 (1): 235–60.
- Hall, Andrew B. 2015. "What Happens When Extremists Win Primaries?" *American Political Science Review* 109 (1): 18–42.
- Hargrave, George E., Deirdre Hiatt, and Tim W. Gaffney. 1988. "F+4+9+Cn: An MMPI Measure of Aggression in Law Enforcement Officers and Applicants." *Journal of Police Science and Administration* 16 (4): 268–73.
- Horan, Kathleen. 2013. "NYPD Memo Directs Officers to Provide Narrative Description for Every Stop and Frisk." *WNYC News*, March 27. <http://www.wnyc.org/story/278670-blog-nypd-memo-directs-officers-provide-narrative-description-every-stop-and-frisk/>.
- Howell, Babe. 2009. "Broken Lives from Broken Windows: The Hidden Costs of Aggressive Order-Maintenance Policing." *New York University Review of Law and Social Change* 33 (1): 271–329.
- Huber, John D., and Charles R. Shipan. 2002. *Deliberate Discretion? The Institutional Foundations of Bureaucratic Autonomy*. Cambridge: Cambridge University Press.
- Imbens, Guido W., and Thomas Lemieux. 2008. "Regression Discontinuity Designs: A Guide to Practice." *Journal of Econometrics* 142 (2): 615–35.
- James, Becca. 2015. "A Look at Stop-and-Frisk Data in Philadelphia, 3 Other Cities." *Philadelphia Inquirer*, March 2.
- Laguna, Louis, Ashley Linn, Kyle Ward, and Rasa Rupslaukyte. 2009. "An Examination of Authoritarian Personality Traits among Police Officers." *Journal of Police and Criminal Psychology* 25 (2): 99–104.
- Legewie, Joscha. 2016. "Racial Profiling and Use of Force in Police Stops: How Local Events Trigger Periods of Increased Discrimination." *American Journal of Sociology* 122 (2) 379–424.
- Lerman, Amy E., and Vesla M. Weaver. 2014a. *Arresting Citizenship: The Democratic Consequences of American Crime Control*. Chicago: University of Chicago Press.
- Lerman, Amy E., and Vesla M. Weaver. 2014b. "Staying Out of Sight? Concentrated Policing and Local Political Action." *Annals of the American Academy of Political and Social Science* 651 (1): 202–19.
- Levine, Deena R., Philip R. Harris, and Herbert Z. Wong. 2002. *Multicultural Law Enforcement: Strategies for Peacekeeping in a Diverse Society*. Upper Saddle River, NJ: Prentice Hall.
- Lipsky, Michael. 1980. *Street-Level Bureaucracy: Dilemmas of the Individual in Public Service*. New York: Russell Sage.
- Lockwood, Daniel, and Ariane Prohaska. 2015. "Police Officer Gender and Attitudes toward Intimate Partner Violence: How Policy Can Eliminate Stereotypes." *International Journal of Criminal Justice Studies* 10 (1): 77–90.
- MacDonald, John, Jeffrey Fagan, and Amanda Geller. 2016. "The Effects of Local Police Surges on Crime and Arrests in New York City." *PLoS ONE* 11 (6): e0157223.
- Martin, Jose. 2014. "Policing Is a Dirty Job, but Nobody's Gotta Do It: 6 Ideas for a Cop-Free World." *Rolling Stone*, December 16.
- McCubbins, Mathew D., Roger G. Noll, and Barry R. Weingast. 1987. "Administrative Procedures as Instruments of Political Control." *Journal of Law, Economics, and Organization* 3 (2): 243–77.
- McNamara, John H. 1967. "Uncertainties in Police Work: The Relevance of Police Recruits? Background and Training." In D. J. Bordua, ed., *The Police: Six Sociological Essays*. New York: Wiley.
- Meredith, Mark, and Michael Morse. 2015. "Discretionary Disenfranchisement: The Case of Legal Financial Obligations." Working paper, University of Pennsylvania.

- Miller, Gary J. 2005. "The Political Evolution of Principal-Agent Models." *Annual Review of Political Science* 8 (1): 203–25.
- Morgan, Stephen L., and Christopher Winship. 2014. *Counterfactuals and Causal Inference*. Cambridge: Cambridge University Press.
- Niederhoffer, Arthur. 1967. *Behind the Shield: The Police in Urban Society*. Garden City, NY: Doubleday.
- Olken, Benjamin A. 2010. "Direct Democracy and Local Public Goods: Evidence from a Field Experiment in Indonesia." *American Political Science Review* 104 (2): 243–67.
- Orne, Martin T. 1962. "On the Social Psychology of the Psychological Experiment: With Particular Reference to Demand Characteristics and Their Implications." *American Psychologist* 17 (11): 776–83.
- Paluck, Elizabeth, and Donald Green. 2009. "Deference, Dissent, and Dispute Resolution: An Experimental Intervention Using Mass Media to Change Norms and Behavior in Rwanda." *American Political Science Review* 103 (4): 622–44.
- Parascandola, Rocco. 2015. "EXCLUSIVE: No Link Found between Sharp Drop in Street Stops, Rise in Shootings Finds Internal NYPD Investigation." *New York Daily News*, March 7. <http://www.nydailynews.com/new-york/nyc-crime/exclusive-no-link-stop-drop-shooting-rise-nypd-article-1.2140591>.
- Parascandola, Rocco, Kerry Burke, and Larry McShane. 2015. "EXCLUSIVE: Huge Drop in Stop-and-Frisk as NYC Crime Increases Raises Fear That Cops are Reluctant to Confront Criminals." *New York Daily News*, June 5.
- Parascandola, Rocco, Jenna O'Donnell, and Larry McShane. 2014. "EXCLUSIVE: NYPD Stop-and-Frisks Drop 99% in Brooklyn, While Shootings Increase in Brownsville, East New York." *New York Daily News*, August 16. <http://www.nydailynews.com/new-york/nyc-crime/nypd-stop-and-frisks-drop-99-percent-shootings-increase-brooklyn-article-1.1905456>.
- Persico, Nicola, and David A. Castleman. 2005. "Detecting Bias: Using Statistical Evidence to Establish Intentional Discrimination in Racial Profiling Cases." *University of Chicago Legal Forum* 1:217–35.
- 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." *Economic Journal* 116: F351–F367.
- Rayman, Graham A. 2013. *The NYPD Tapes: A Shocking Story of Cops, Cover-Ups, and Courage*. London: Macmillan.
- Roberg, Robert R. 1978. "Analysis of the Relationships among Higher Education Belief Systems and Job Performance of Patrol Officers." *Journal of Police Science and Administration* 6 (3): 336–44.
- Rosenfeld, Richard, and Robert Fornago. 2012. "The Impact of Police Stops on Precinct Robbery and Burglary Rates in New York City, 2003–2010." *Justice Quarterly* 31 (1): 96–122.
- Rubén, Hernández-Murillo, and John Knowles. 2004. "Racial Profiling or Racist Policing? Bounds Tests in Aggregate Data." *International Economic Review* 45:959–87.
- Sampson, Robert J., and Charles Loeffler. 2010. "Punishment's Place: The Local Concentration of Mass Incarceration." *Daedalus* 139 (3): 20–31.
- Schmidt, Michael S. 2015. "Scant Data Frustrates Efforts to Assess Number of Shootings by Police." *New York Times*, April 8.
- Sellbom, Martin, Gary L. Fischler, and Yossef S. Ben-Porath. 2007. "Identifying MMPI-2 Predictors of Police Officer Integrity and Misconduct." *Criminal Justice and Behavior* 35 (1): 985–1004.
- Shadish, William R., Thomas D. Cook, and Donald Thomas Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Wadsworth Cengage Learning.
- Skolnick, Jerome H. 1977. "A Sketch of the Policeman's 'Working Personality.'" In Daniel B. Kennedy, ed., *The Dysfunctional Alliance: Emotion and Reason in Justice Administration*. Cincinnati: Anderson.
- Smith, Brad. 2003. "The Impact of Police Officer Diversity on Police-Caused Homicides." *Policy Studies* 31 (2): 147–62.
- Travis, Jeremy, Bruce Western, and Steve Redburn, eds. 2014. *The Growth of Incarceration in the United States: Exploring Causes and Consequences*. Washington, DC: National Academies.
- Twersky-Glasner, Aviva. 2005. "Police Personality: What Is It and Why Are They like That?" *Journal of Police and Criminal Psychology* 20 (1): 56–67.
- Tyler, Tom R., and Jeffrey Fagan. 2008. "Why Do People Cooperate with the Police?" *Ohio State Journal of Criminal Law* 6 (1): 231–75.
- Tyler, Tom R., and Cheryl J. Wakslak. 2004. "Profiling and Legitimacy of the Police: Procedural Justice, Attributions of Motive, and the Acceptance of Social Authority." *Criminology* 42 (1): 13–42.
- Vitale, Alex S. 2014. "We Don't Just Need Nicer Cops: We Need Fewer Cops." *The Nation*, December 4.
- Wilson, James Q. 1968. *Varieties of Police Behavior*. Cambridge, MA: Harvard University Press.
- Wilson, James Q., and George L. Kelling. 1982. "Broken Windows: The Police and Neighborhood Safety." *Atlantic Monthly* 249 (3): 29–38.

**Online Appendix: “Modern Police Tactics,
Police-Citizen Interactions and the Prospects for
Reform”**

Appendix A: Notes on Data Cleaning and Model Specifications

Data Cleaning

Cleaned and merged SQF data up to 2013 were generously supplied by the authors of Goel et al. (2016). The authors made several reasonable alterations, (maintained here), including dropping cases where the suspect's age does not fall between 10 and 80, as these are likely mis-codings. Additionally, cases in which the year of the stop was listed as "1900" (47 observations) were dropped for the same reason. Weapon stops where the outcome of the weapon stop was not recorded were also dropped from all analyses. Using code from Goel et al. (2016) as a guide, I also obtained raw data from 2014-2015 from the NYPD web site, processed it, and appended it to the earlier data. Following Goel et al. (2016), the suspected crime indicator was coded using the "detailCM" field in the raw SQF data.

Note: the weekly crime statistics supplied by the NYPD used in the analysis of homicides show 532, 445, 349 and 295 homicides in the years 2010, 2011, 2012 and 2013, respectively. These are lower figures than the final annual murder totals reported by the NYPD, (which total 536, 515, 419 and 335 for the same years), likely because the weekly crime reports are preliminary in nature and exclude crimes discovered at later dates.

Model Specifications

The treatment effects in the main manuscript were all estimated via ordinary least squares regression. As noted on page ?? in the main text, $s_j(d_i)$ is specified as either a linear, quadratic or cubic function, with d_i denoting the distance in days from the intervention. Below are details on these model specifications. In each model, the parameter of interest is τ , the immediate change in the hit rate the day of the intervention.

Linear:

$$weapon_i = \alpha + \tau memo_i + \beta_1 d_i + \beta_2 memo_i * d_i + \epsilon_i$$

The quadratic model was specified as follows:

Quadratic:

$$\begin{aligned} weapon_i = & \alpha + \tau memo_i + \beta_1 d_i + \beta_2 memo_i * d_i \\ & + \beta_3 d_i^2 + \beta_4 memo_i * d_i^2 + \epsilon_i \end{aligned}$$

The cubic model was specified as follows:

Cubic:

$$\begin{aligned} weapon_i = & \alpha + \tau memo_i + \beta_1 d_i + \beta_2 memo_i * d_i \\ & + \beta_3 d_i^2 + \beta_4 memo_i * d_i^2 + \beta_5 d_i^3 + \beta_6 memo_i * d_i^3 + \epsilon_i \end{aligned}$$

Figure A1: A picture of Hall's memo, distributed to all patrol units on March 5, 2013.

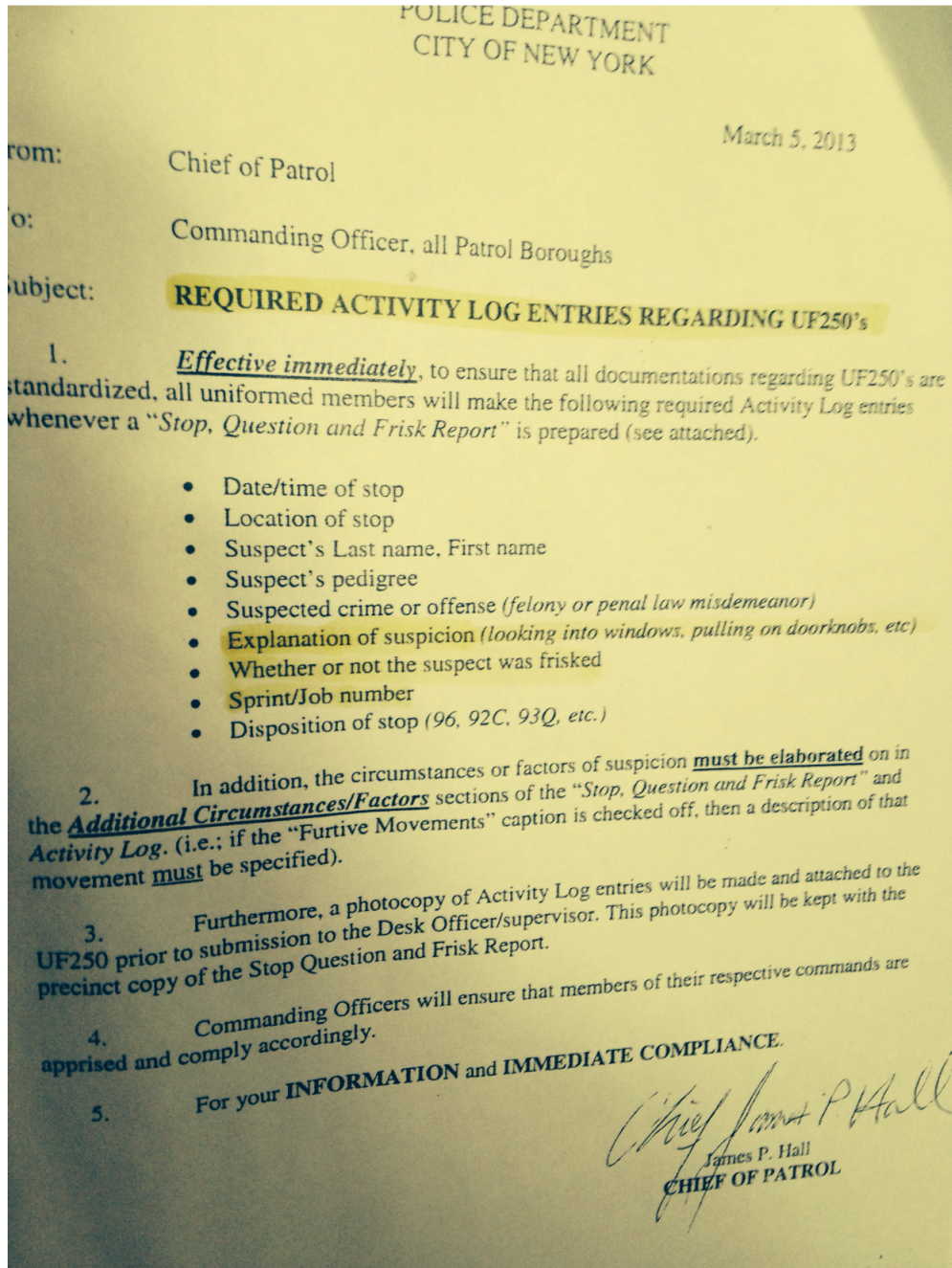



Figure A2: The front side of a UF-250 form, used by officers to record the circumstances and outcomes of each stop of a criminal suspect. Source: NYCLU.org.

(COMPLETE ALL CAPTIONS)

 STOP, QUESTION AND FRISK REPORT WORKSHEET PD344-151A (Rev. 11-02)		Pct. Serial No.	
		Date	Pct. Of Occ.
Time Of Stop	Period Of Observation Prior To Stop	Radio Run/Sprint #	
Address/Intersection Or Cross Streets Of Stop			
<input type="checkbox"/> Inside	<input type="checkbox"/> Transit	Type Of Location	
<input type="checkbox"/> Outside	<input type="checkbox"/> Housing	Describe:	
Specify Which Felony/P.L. Misdemeanor Suspected			Duration Of Stop
What Were Circumstances Which Led To Stop? (MUST CHECK AT LEAST ONE BOX)			
<input type="checkbox"/> Carrying Objects In Plain View Used In Commission Of Crime e.g., Slim Jim/Pry Bar, etc.		<input type="checkbox"/> Actions Indicative Of Engaging In Drug Transaction.	
<input type="checkbox"/> Fits Description.		<input type="checkbox"/> Furtive Movements.	
<input type="checkbox"/> Actions Indicative Of "Casing" Victim Or Location.		<input type="checkbox"/> Actions Indicative Of Engaging In Violent Crimes.	
<input type="checkbox"/> Actions Indicative Of Acting As A Lookout.		<input type="checkbox"/> Wearing Clothes/Disguises Commonly Used In Commission Of Crime.	
<input type="checkbox"/> Suspicious Bulge/Object (Describe)			
<input type="checkbox"/> Other Reasonable Suspicion Of Criminal Activity (Specify)			
Name Of Person Stopped		Nickname/ Street Name	Date Of Birth
Address		Apt. No.	Tel. No.
Identification: <input type="checkbox"/> Verbal <input type="checkbox"/> Photo I.D. <input type="checkbox"/> Refused			
<input type="checkbox"/> Other (Specify)			
Sex: <input type="checkbox"/> Male Race: <input type="checkbox"/> White <input type="checkbox"/> Black <input type="checkbox"/> White Hispanic <input type="checkbox"/> Black Hispanic			
<input type="checkbox"/> Female <input type="checkbox"/> Asian/Pacific Islander <input type="checkbox"/> American Indian/Alaskan Native			
Age	Height	Weight	Hair Eyes Build
Other (Scars, Tattoos, Etc.)			
Did Officer Explain Reason For Stop		If No, Explain:	
<input type="checkbox"/> Yes <input type="checkbox"/> No			
Were Other Persons Stopped/ Questioned/Frisked?		<input type="checkbox"/> Yes	If Yes, List Pct. Serial Nos.
		<input type="checkbox"/> No	
If Physical Force Was Used, Indicate Type:			
<input type="checkbox"/> Hands On Suspect		<input type="checkbox"/> Drawing Firearm	
<input type="checkbox"/> Suspect On Ground		<input type="checkbox"/> Baton	
<input type="checkbox"/> Pointing Firearm At Suspect		<input type="checkbox"/> Pepper Spray	
<input type="checkbox"/> Handcuffing Suspect		<input type="checkbox"/> Other (Describe)	
<input type="checkbox"/> Suspect Against Wall/Car			
Was Suspect Arrested?	Offense	Arrest No.	
<input type="checkbox"/> Yes <input type="checkbox"/> No			
Was Summons Issued?	Offense	Summons No.	
<input type="checkbox"/> Yes <input type="checkbox"/> No			
Officer In Uniform?	If No, How Identified? <input type="checkbox"/> Shield <input type="checkbox"/> I.D. Card		
<input type="checkbox"/> Yes <input type="checkbox"/> No	<input type="checkbox"/> Verbal		

Results Using Alternate Specifications, Bandwidths

While conditioning on the suspected crime being “criminal possession of a weapon” is optimal for conceptualizing the outcome of interest, (i.e. the degree to which the stop produced evidence of the stated reason for the stop), some may wonder whether conditioning on a stop attribute may impose post-treatment bias (King and Zeng 2006). As Figure A3 shows, there appears to be little threat of such bias in this case, as the rate at which weapons are suspected does not appear to change discontinuously at the treatment boundary.

Further, Table A1 displays the estimated change in the probability of finding a weapon at the intervention among all stops (whether or not a weapon was suspected), and Figure A4 displays the same estimates in narrow bandwidths around the intervention. Among all stops, the hit rate during the pre-treatment period was substantially smaller than among weapon stops, at 1.2% vs. 3.5%, respectively. This makes sense, since there should be a lower probability of finding a weapon when a weapon is not suspected. The estimated treatment effects, too, are smaller among all stops, but still represent substantial proportional increases. Using the full data set, the largest effect size is 1.2 percentage points, a doubling of the hit rate. Comparable effects are estimated in narrow bandwidths (see Figure A4), though given the smaller size of the discontinuity, it is more difficult to discern these estimates from zero when subsetting to these small windows of data.

Table A1: OLS Estimates of Discontinuity, All Stops 2008-2015

	Diff. in Means	Diff. in Means [†]	Linear	Linear [†]	Quadratic	Quadratic [†]	Cubic	Cubic [†]
$\hat{\tau}$	0.017* (0.000)	0.009* (0.001)	0.008* (0.001)	0.005* (0.001)	0.008* (0.001)	0.003* (0.001)	0.003* (0.001)	0.001 (0.001)
N	3,184,857	3,183,950	3,184,857	3,183,950	3,184,857	3,183,950	3,184,857	3,183,950

[†] Includes controls for year, month, day of week, and prior day’s hit rate.

Maximum of homoscedastic and HAC standard errors in parentheses. * indicates $p < .05$, two-tailed.

Figure A3

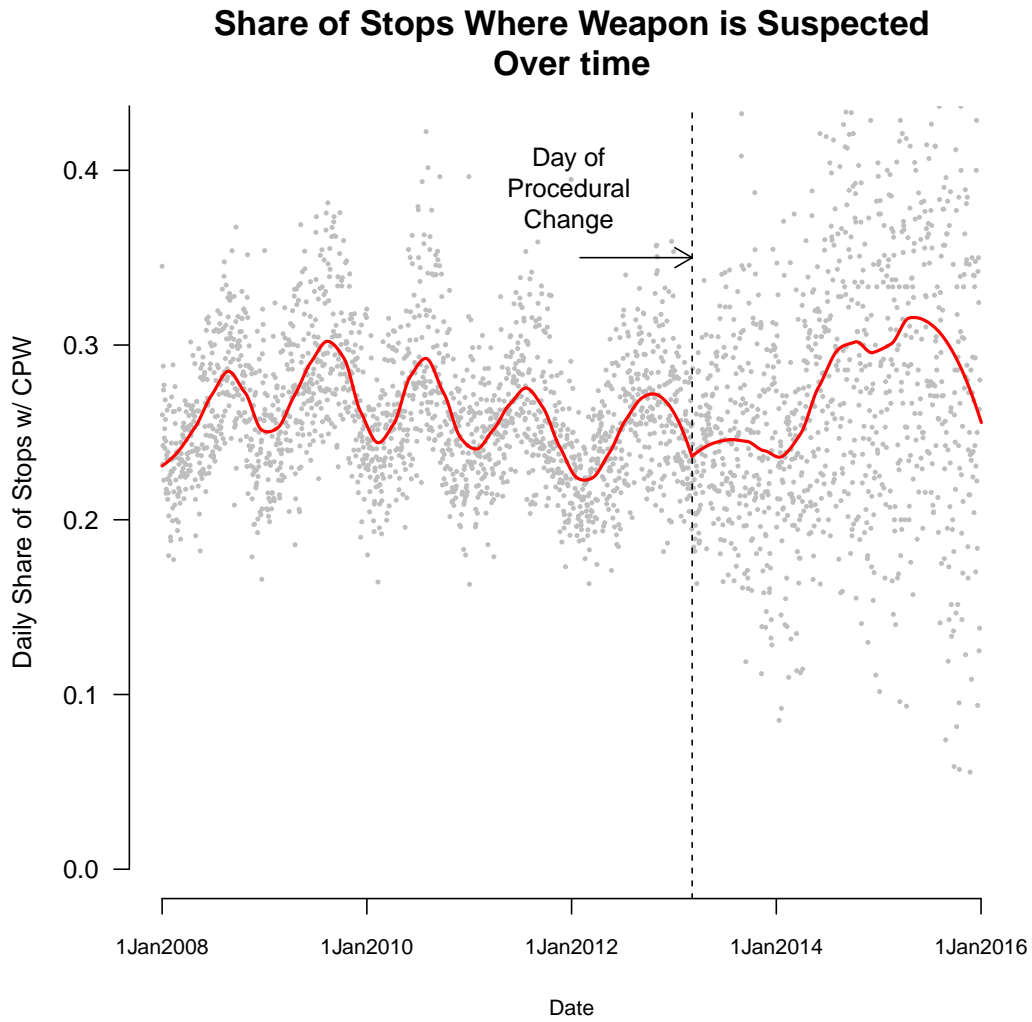
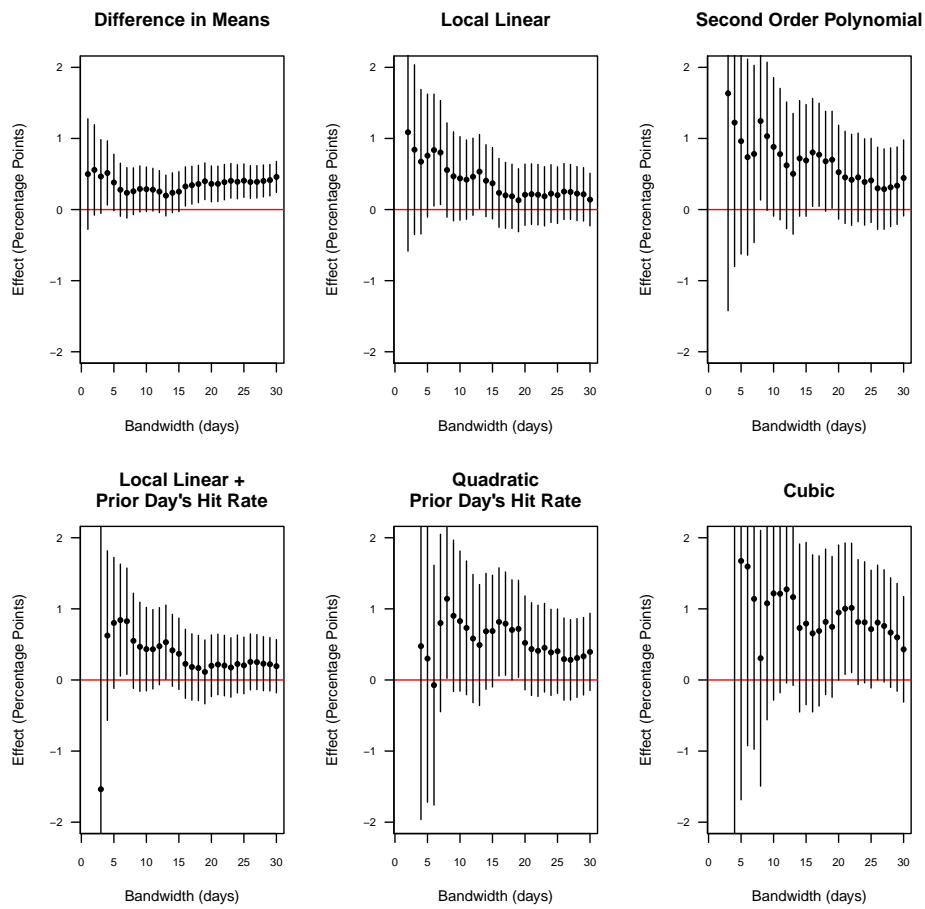


Figure A4: Estimated change in the weapon recovery rate **among all stops**.



Alternate bandwidths for numerator/denominator analysis

When using the full range of daily sums during 2008-2015, there remains robust evidence that the number of weapon stops conducted declined sharply the day of the intervention (the hit rate's denominator). There is no consistent evidence that number of stops producing a weapon (the hit rate's numerator) increased, (the coefficients are not consistently signed).

Table A2: OLS Estimates of Discontinuity in Number of Stops Producing a Weapon (Numerator), Weapon Stops 2008-2015

	Diff. in Means	Diff. in Means [†]	Linear	Linear [†]	Quadratic	Quadratic [†]	Cubic	Cubic [†]
$\hat{\tau}$	-10.578* (0.441)	-3.315* (0.665)	-6.523* (0.693)	-3.465 (0.682)	-0.392 (0.756)	-0.487 (0.759)	3.111* (0.911)	1.055 (0.857)
N	2,921	2,920	2,921	2,920	2,921	2,920	2,921	2,920

Table A3: OLS Estimates of Discontinuity in Number of Weapon Stops Conducted (Denominator), Weapon Stops 2008-2015

	Diff. in Means	Diff. in Means [†]	Linear	Linear [†]	Quadratic	Quadratic [†]	Cubic	Cubic [†]
$\hat{\tau}$	-366.831* (16.985)	-80.111* (10.382)	-284.276* (21.308)	-88.45* (10.726)	-128.93* (20.083)	-61.075* (11.76)	-29.684 (21.539)	-37.998* (13.169)
N	2,921	2,920	2,921	2,920	2,921	2,920	2,921	2,920

Results using an optimum bandwidth

The results in the main text are generated using an array of bandwidths in order to demonstrate the robustness of treatment effects to specification choices. We can also compute the treatment effect using a technique to derive an optimum bandwidth given the data (Imbens and Kalyanaraman 2011; Calonico, Cattaneo and Titunik 2014). Using the `rdrobust` function (Calonico, Cattaneo and Titunik 2014) to estimate the immediate increase in the weapon recovery rate among weapon stops the day of the intervention (given an estimated optimum bandwidth of about 186 days) produces a point estimate of 1.69 percentage points

(robust SE = 0.376; bias-corrected SE = 0.339), which is highly comparable to the core results in the main text.

Alternate standard errors

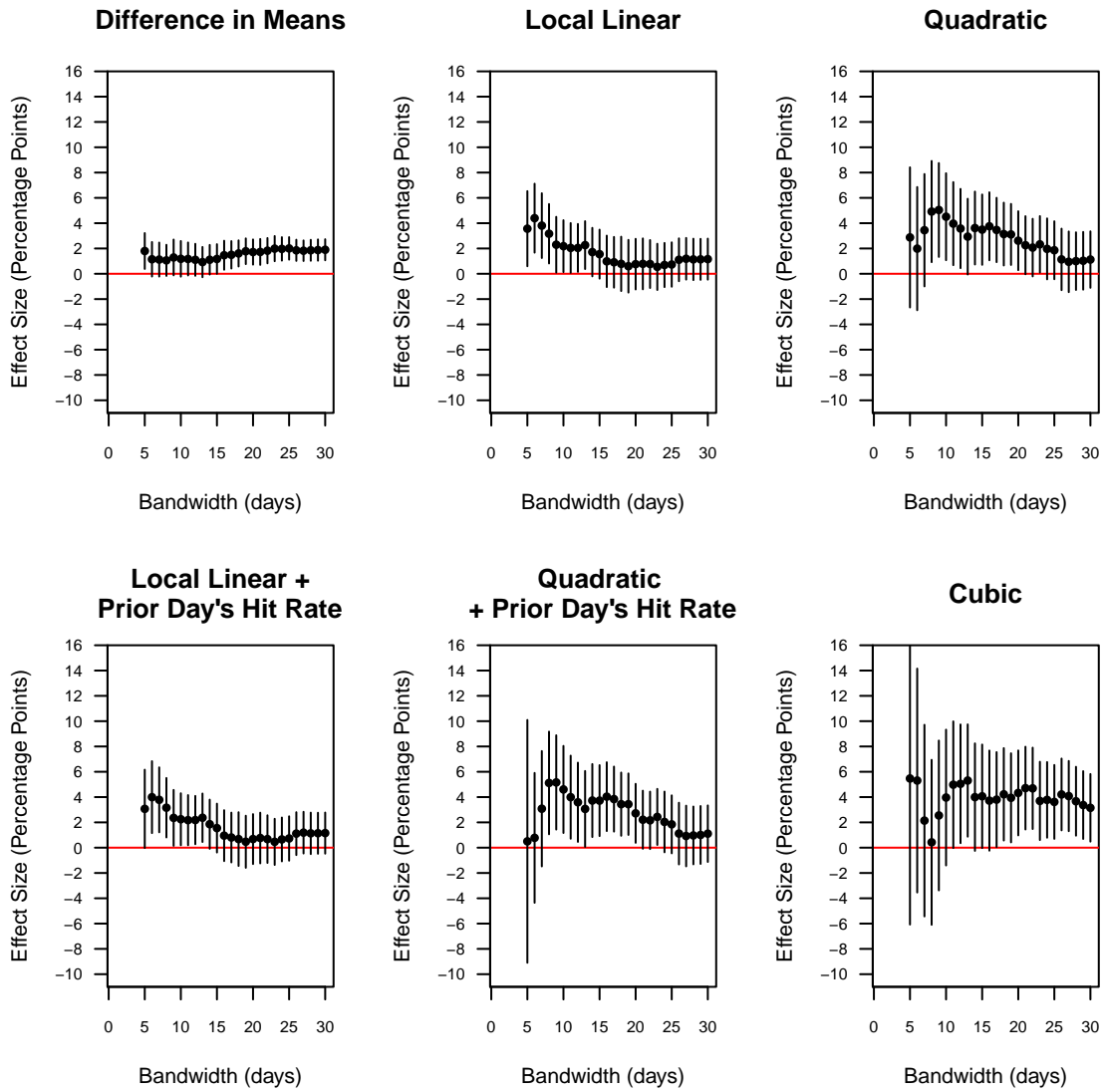
Table A4: Treatment effects with standard errors clustered by precinct

	Diff. in Means	Diff. in Means [†]	Linear	Linear [†]	Quadratic	Quadratic [†]	Cubic	Cubic [†]
$\hat{\tau}$	0.0514 * (0.0073)	0.0315 * (0.0047)	0.0304 * (0.0054)	0.0216 * (0.0046)	0.0287 * (0.0046)	0.0196 * (0.0041)	0.0132 * (0.0048)	0.0105 * (0.0042)
N	826,573	826,260	826,573	826,260	826,573	826,260	826,573	826,260

Standard errors clustered by precinct in parentheses.

* indicates significance at $p < 0.05$

Figure A5: Estimated change in the weapon recovery rate in local bandwidths with standard errors clustered by precinct.



Appendix B: Other Measures of Increased Quality

As perviously mentioned, whether a weapon was discovered after a weapon is suspected is the cleanest available measure of whether a stop was justified, since it links the outcome of the stop to the specific reason the stop was conducted. Other versions of the hit rate lack this feature—i.e., it is unclear whether they have the correct denominator—making their substantive interpretation more difficult. This limitation aside, we may wish to estimate the degree to which other outcomes of stops changed with the intervention.

Tables B1-B3 display the estimated discontinuities in rates of arrests, finding contraband and issuing summonses, respectively, using all data points between 2008 and 2013 (not just weapon stops). The rate of arrests, an indication of more serious—and perhaps more readily observable—offenses, jumps discontinuously by a large amount in most specifications (between 1 and 5 percentage points), at the moment of the intervention. The rate at which contraband was discovered appears to have increased by smaller amounts (less than one percentage point), and the rate at which summonses were issued showed no consistent change across specifications.

Figures B1-B3 display the results when estimating these discontinuities in narrow temporal windows (30 days or less on either side of the intervention). When subjected to this more conservative test, it appears that the arrest rate increases markedly with the intervention across most bandwidths, while there is weak evidence for a jump in the rate at which contraband was found, and little evidence of an effect on issuing summonses.

Figure B1: Local estimates of the discontinuity in the **arrest rate**

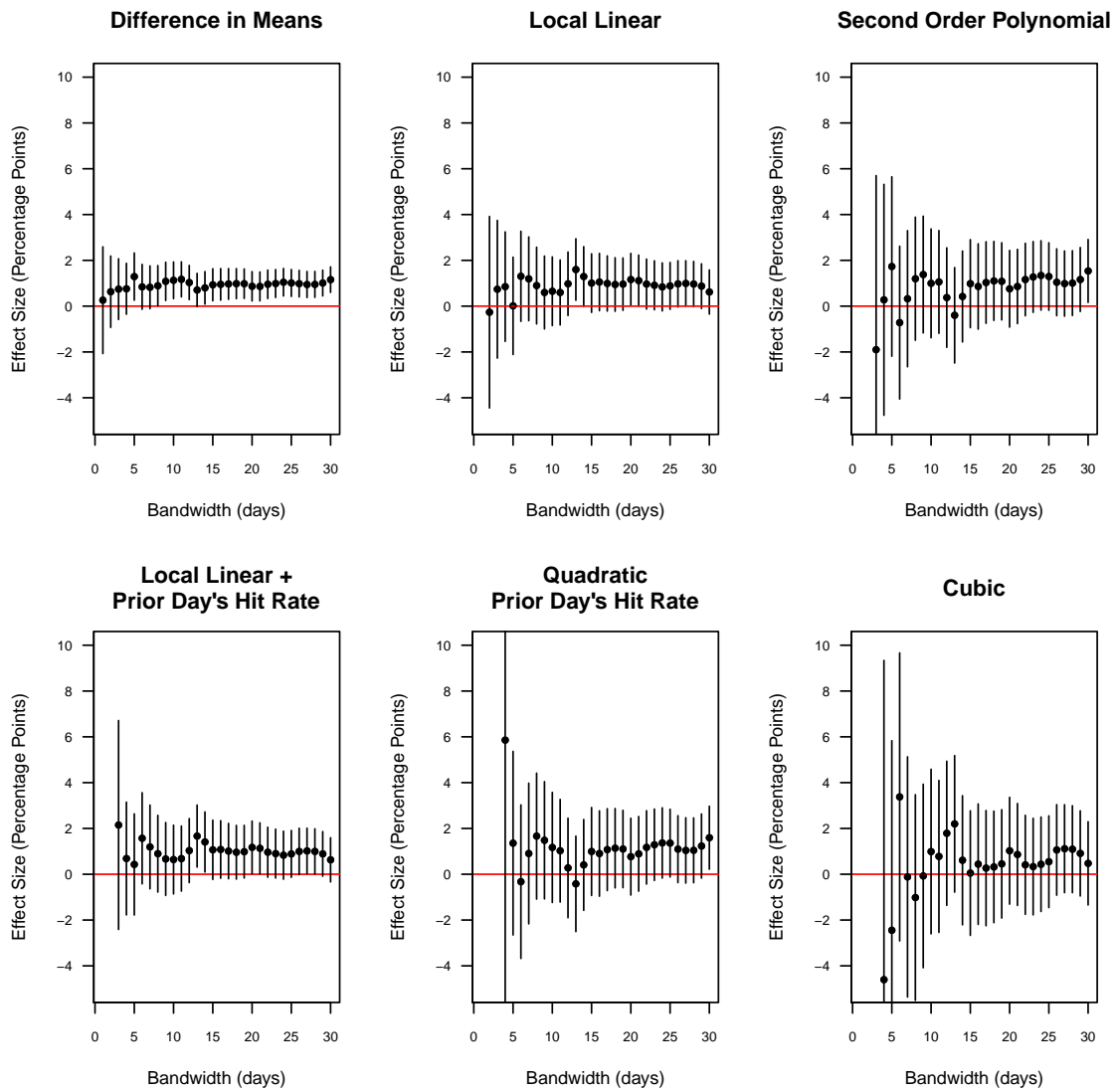


Figure B2: Local estimates of the discontinuity in the rate of **finding contraband**.

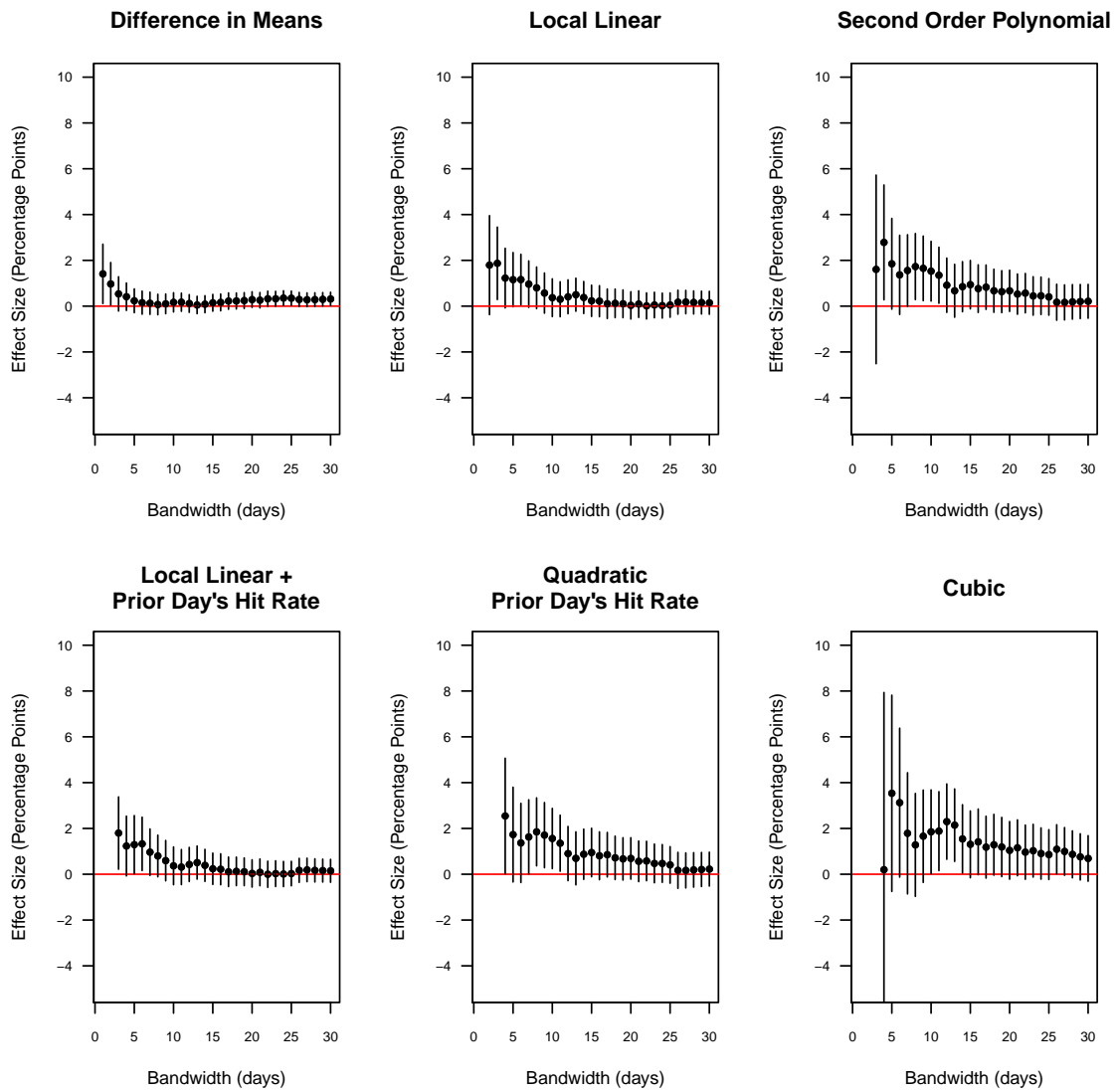


Figure B3: Local estimates of the discontinuity in the **summons rate**.

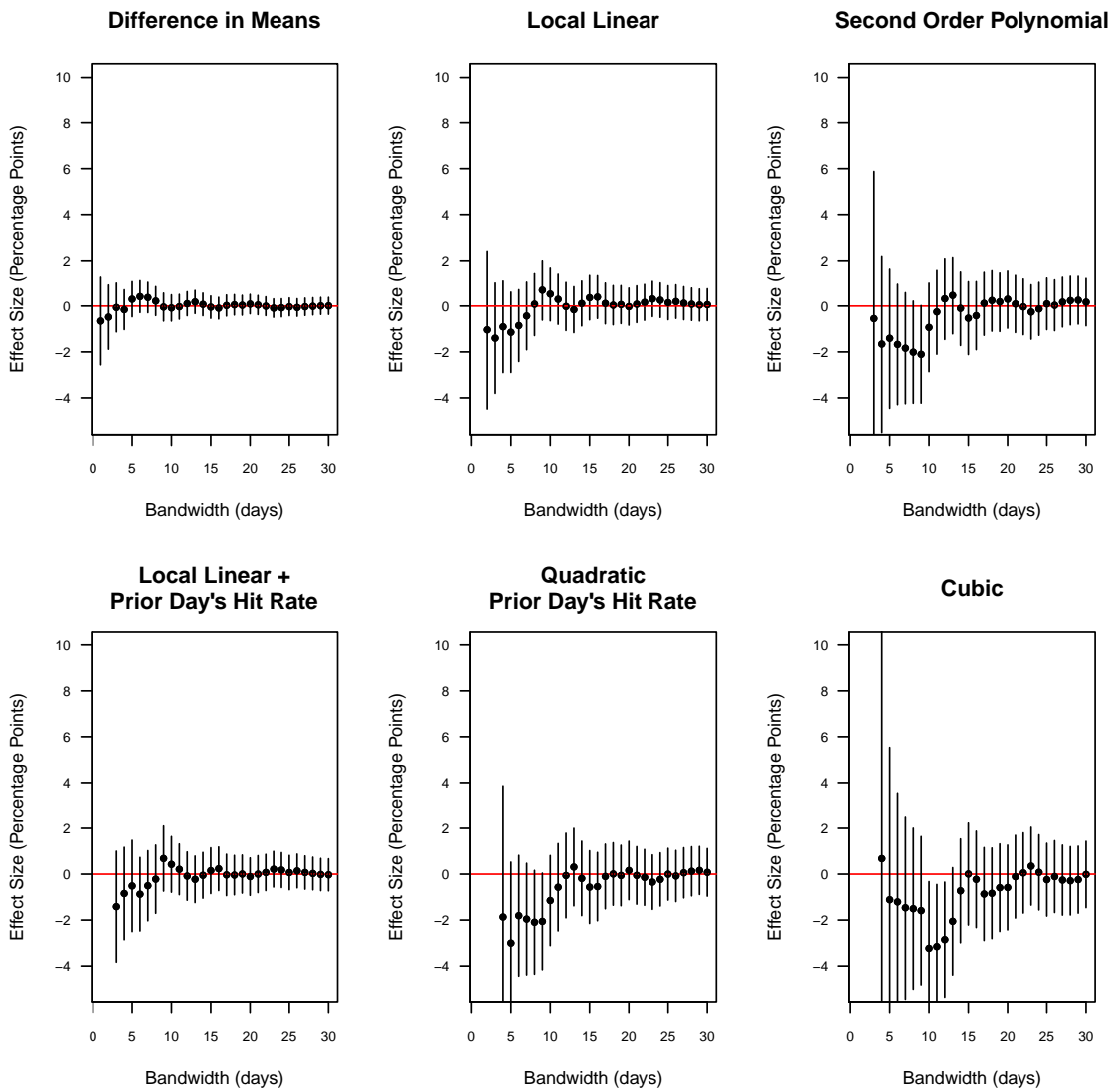


Table B1: OLS Estimates of Discontinuity, Arrest Rate, All Stops

	Diff. in Means	Diff. in Means [†]	Linear	Linear [†]	Quadratic	Quadratic [†]	Cubic	Cubic [†]
$\hat{\tau}$	0.057* (0.001)	0.019* (0.002)	0.020* (0.001)	0.008* (0.002)	0.012* (0.002)	0.002 (0.002)	-0.006* (0.002)	-0.008* (0.003)
N	3,184,857	3,183,950	3,184,857	3,183,950	3,184,857	3,183,950	3,184,857	3,183,950

[†] Includes controls for year, month, day of week, and prior day's hit rate.

Maximum of homoscedastic and HAC standard errors in parentheses. * indicates $p < .05$, two-tailed.

Table B2: OLS Estimates of Discontinuity in Rate of Discovering Contraband, All Stops

	Diff. in Means	Diff. in Means [†]	Linear	Linear [†]	Quadratic	Quadratic [†]	Cubic	Cubic [†]
$\hat{\tau}$	0.015* (0.001)	0.007* (0.001)	0.006* (0.001)	0.003* (0.001)	0.005* (0.001)	0.003* (0.001)	0.001 (0.001)	-0.001 (0.001)
N	3,184,730	3,183,823	3,184,730	3,183,823	3,184,730	3,183,823	3,184,730	3,183,823

[†] Includes controls for year, month, day of week, and prior day's hit rate.

Maximum of homoscedastic and HAC standard errors in parentheses. * indicates $p < .05$, two-tailed.

Table B3: OLS Estimates of Discontinuity, Rate of Issuing Summonses, All Stops

	Diff. in Means	Diff. in Means [†]	Linear	Linear [†]	Quadratic	Quadratic [†]	Cubic	Cubic [†]
$\hat{\tau}$	-0.028* (0.001)	-0.003* (0.001)	-0.014* (0.001)	-0.004* (0.001)	0.000 (0.001)	-0.003* (0.001)	0.008* (0.001)	-0.001 (0.002)
N	3,184,857	3,183,950	3,184,857	3,183,950	3,184,857	3,183,950	3,184,857	3,183,950

[†] Includes controls for year, month, day of week, and prior day's hit rate.

HAC standard errors in parentheses. * indicates $p < .05$, two-tailed.

Table B4: Estimated Discontinuity in Weekly Homicides on Week of March 5, 2013

	Difference in Means	Difference in Means [†]	Linear	Linear [†]	Second Order Polynomial	Second Order Polynomial [†]	Cubic	Cubic [†]
Change at Threshold	-2.64 *	-3.27 *	-0.61	-0.92	-0.94	-0.13	-1.83	-0.55
	(0.620)	(0.60)	(1.16)	(1.23)	(1.69)	(1.80)	(2.20)	(2.37)
<i>N</i>	208	208	208	208	208	208	208	208

[†] Includes month fixed effects.

Homoscedastic standard errors in parentheses. * indicates $p < 0.05$, two-tailed.

Figure B4: **Frequency of robberies over time:** The figure shows the weekly robberies over time. There is no evidence of an increase at the moment of the intervention, especially once the data are adjusted for monthly seasonality.

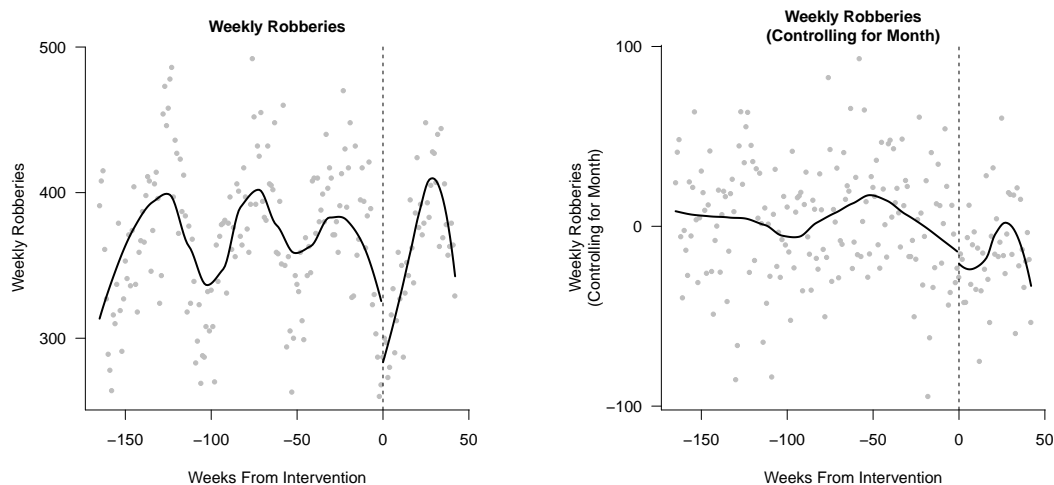


Table B5: Estimated Discontinuity in Weekly Robberies on Week of March 5, 2013

	Difference in Means	Difference in Means [†]	Linear	Linear [†]	Quadratic	Quadratic [†]	Cubic	Cubic [†]
Change at Threshold	-8.78	-18.16 *	-68.14 *	-26.32 *	-99.03 *	-41.26 *	-65.64 *	-2.10
	(8.46)	(5.70)	(16.05)	(12.58)	(22.86)	(18.24)	(29.56)	(23.58)
<i>N</i>	208	208	208	208	208	208	208	208

[†] Includes month fixed effects.

Homoscedastic standard errors in parentheses. * indicates $p < 0.05$, two-tailed.

Figure B5: **Stop attributes over time - all stops**: Using all stops, the figure displays how the prevalence of various stop attributes changed over time. The intervention appears to have caused the share of suspects who are white to increase, a change offset by a decline in the share of stops who made up by black suspects. This pattern supports to claim that “unnecessary” stops were being abandoned post-treatment, since prior work showed that black suspects were stopped unnecessarily more often than white suspects (Goel et al. 2016). The treatment also appears to have caused the rate of stops justified by observing “furtive movements” to decline. This justification was often derided by critics as arbitrary and vague. Its decline is another indication of officers avoiding stops with weaker legal justification post-treatment.

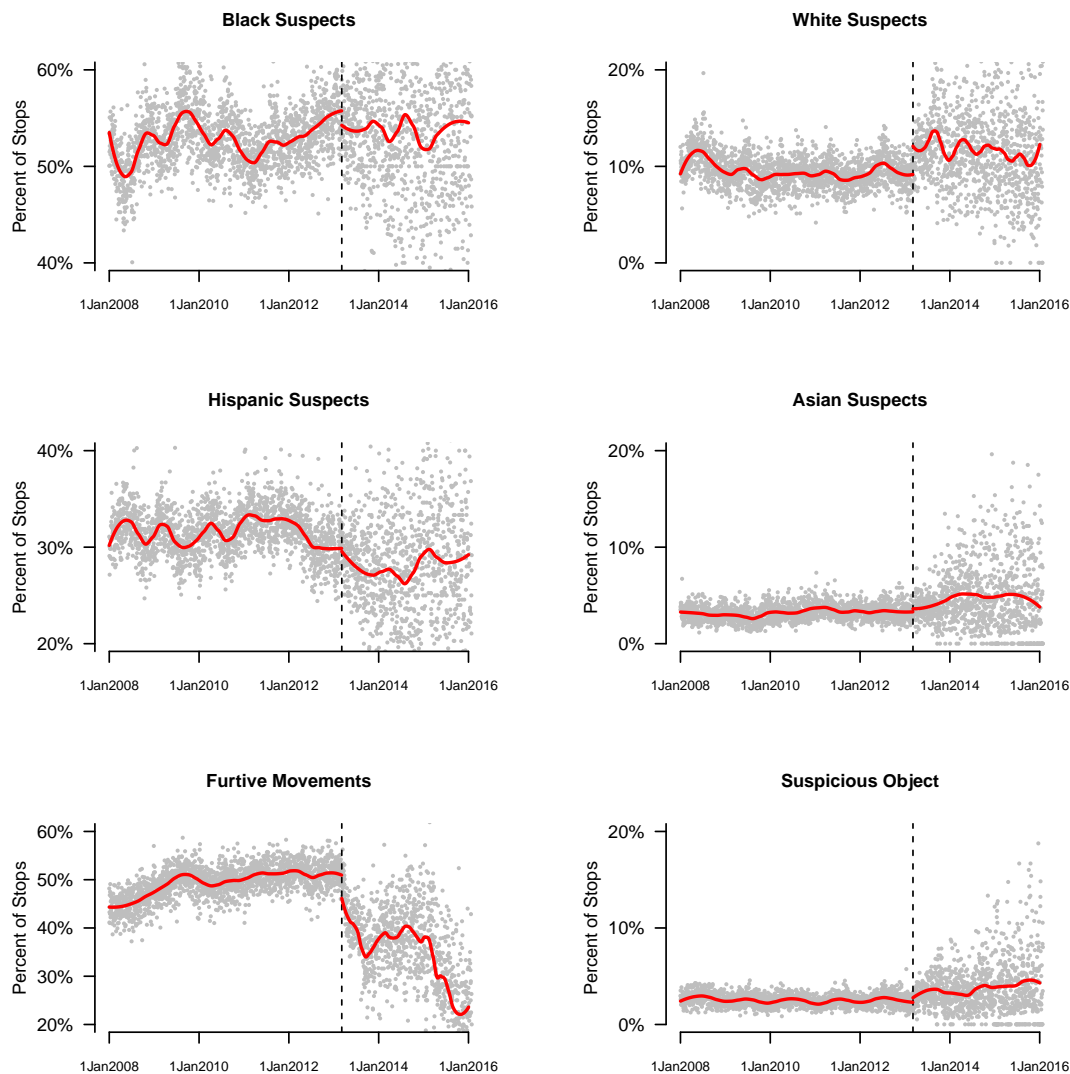
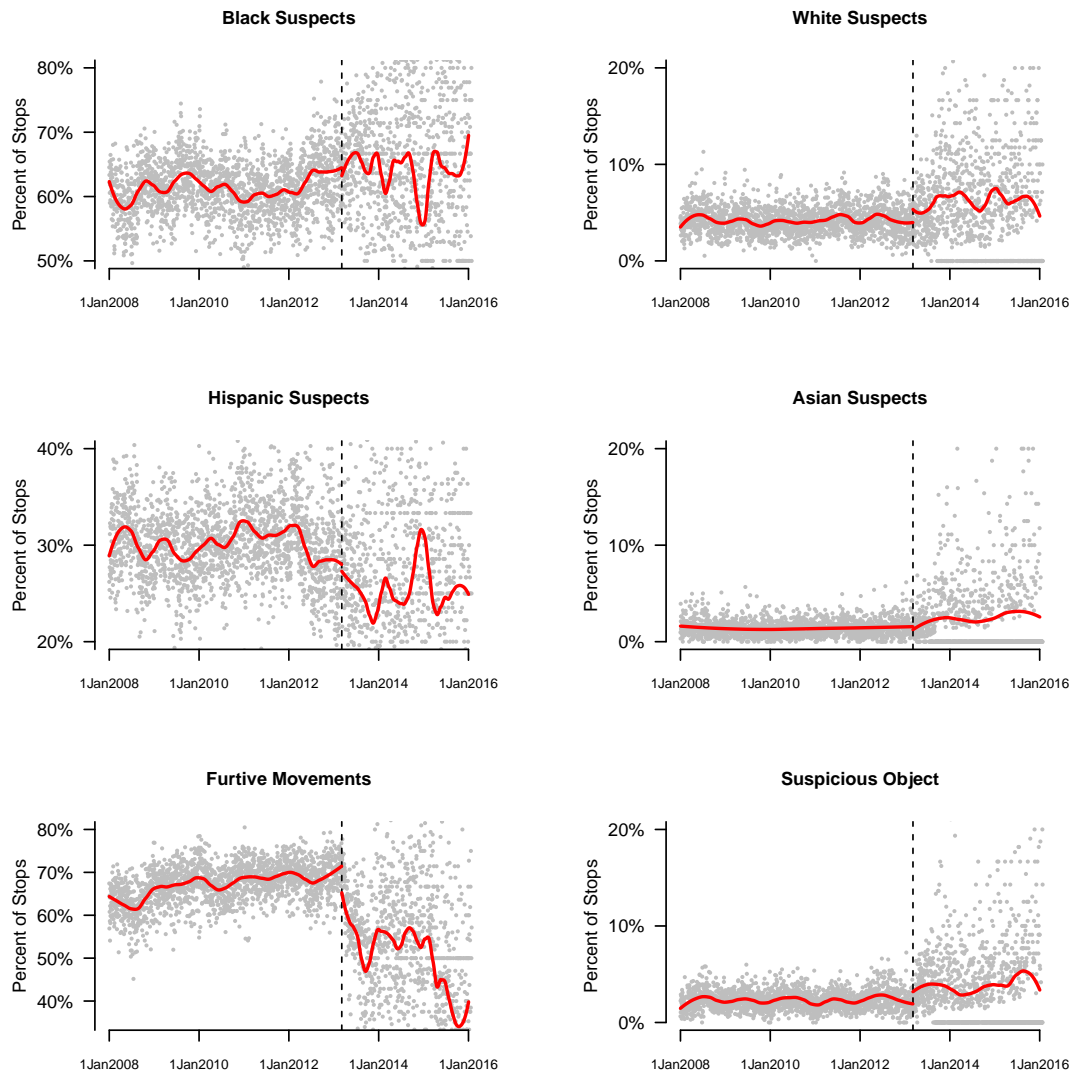


Figure B6: **Stop attributes over time - weapon stops:** Using weapon stops only, the panel displays how the prevalence of various stop attributes changed over time. The intervention appears to have caused the share of suspects who are white to modestly increase, though this result is more difficult to detect than in the previous figure using all stops. The treatment also appears to have caused the rate of stops justified by observing “furtive movements” to decline and may have caused a slight increase in the share of stops where a “suspicious object” was noticed—the latter being a strong predictor of finding a weapon. The “furtive movements” justification was often derided by critics as arbitrary and vague. Its decline is another indication of officers avoiding stops with weaker legal justification post-treatment.



Appendix C: Heterogenous Treatment Effects

To test for geographic heterogeneity in treatment effects, I placed pre-treatment observations in the SQF data in Census block groups using the longitude and latitude markers provided by the NYPD, and then matched the unique Census block groups in the SQF data to Census demographic data from 2010. Ninety-nine percent of SQF observations were successfully paired with block group data. Following Hainmueller, Mummolo and Xu (2016) I discretized all moderators into high and low bins, to avoid the pitfalls of interacting the treatment with a continuous variable while preserving as much statistical power as possible.

For tests involving the racial makeup of block groups, observations were coded as being in the “high” white cell if their block group was at or above the median value of % white among unique block groups in New York City. I also coded stops as being in high or low homicide precincts by computing the homicide per capita rate in each precinct using the mean number of homicides per precinct between 2008-2012 according to publicly available NYPD data¹ and precinct population data generously shared by the authors of Rosenfeld and Fornago (2012).² “High” homicide precincts were those that fell at or above the median for homicides per capita among all unique precincts in New York City. Note: most stops occurred among nonwhite suspects and in places with high shares of nonwhite residents, it

¹ http://www.nyc.gov/html/nypd/downloads/pdf/analysis_and_planning/seven_major_felony_offenses_by_precinct_2000_2015.pdf

² Note, Rosenfeld and Fornago (2012) determined precinct populations via a crosswalk of Census tracts to NYPD precincts. The crosswalk assumed that tract populations were evenly distributed across tract geography. Where tracts crossed into multiple precincts, the authors apportioned the population of the tract into the different precincts based on the proportion of tract geography within each precinct. These data exclude population figures for the precinct covering Central Park, so that precinct is omitted from the precinct-level analysis below.

is difficult to construct well-powered tests of differences in treatment effects between these groups.

In line with the discussion in the main text, Figure C1 and Table C1 show some evidence that treatment effects were larger in block groups with higher shares of white residents. Figure C2 and Table C2, however, show mixed results with regard to disparities in the treatment effect across high and low homicide precincts. Using all available data, there is some indication that the treatment was more effective for white suspects than nonwhite suspects (see Table C3). But using narrow bandwidths, (i.e. tests less prone to omitted variable bias) to conduct the same tests, we recover inconsistently signed point estimates that are imprecisely estimated, making it difficult to draw firm conclusions.

Goel et al. (2016) and others have noted that white suspects enjoy a higher hit rate than non-white suspects, an indication that police may have a higher standard of suspicion for stopping whites than nonwhites. For the intervention to have diminished this gap, the differences in treatment effects in Table C3 and in Figure C3 would have to be negative. Because they appear positive or indiscernible from zero, there is no evidence the intervention erased these disparities. And while there is modest evidence of larger treatment effects among white suspects, which make up only about 10% of stops in this period, it is difficult to infer racial discrimination given this research design, as the race of suspects likely correlates with many unobserved factors that influence the probability of recovering a weapon.

Table C1: OLS Estimates of discontinuity in Census block groups with low % white, high % white, and the difference between the two. All weapon stops 2008-2015.

	Difference in Means	Difference in Means [†]	Linear	Linear [†]	Quadratic	Quadratic [†]	Cubic	Cubic [†]
low % white	0.041* (0.002)	0.024* (0.003)	0.022* (0.002)	0.015* (0.003)	0.025* (0.003)	0.016* (0.004)	0.010* (0.003)	0.007 (0.004)
high % white	0.063* (0.005)	0.044* (0.006)	0.049* (0.008)	0.042* (0.009)	0.048* (0.01)	0.039* (0.011)	0.035* (0.013)	0.032* (0.014)
Diff. in effects	0.022* (0.006)	0.020* (0.006)	0.027* (0.009)	0.027* (0.009)	0.023* (0.011)	0.022* (0.011)	0.025 (0.014)	0.024 (0.014)
<i>N</i>	797,320	797,018	797,320	797,018	797,320	797,018	797,320	797,018

[†] Includes controls for year, month, day of week, and prior day's hit rate.

Maximum of homoscedastic and HAC standard errors in parentheses. * indicates $p < .05$, two-tailed.

Table C2: OLS Estimates of discontinuity in hit rate among precincts with low homicide rates, high homicide rates, and the difference between the two. All weapon stops 2008-2015.

	Difference in Means	Difference in Means [†]	Linear	Linear [†]	Second Order Polynomial	Second Order Polynomial [†]	Cubic	Cubic [†]
low homicide	0.079* (0.005)	0.059* (0.005)	0.047* (0.007)	0.038* (0.007)	0.028* (0.009)	0.018* (0.009)	0.023* (0.011)	0.019 (0.011)
high homicide	0.041* (0.002)	0.023* (0.003)	0.025* (0.002)	0.015* (0.003)	0.026* (0.003)	0.017* (0.003)	0.01* (0.003)	0.007 (0.004)
Difference in effects	-0.038* (0.005)	-0.036* (0.005)	-0.023* (0.007)	-0.023* (0.007)	-0.002 (0.009)	-0.002 (0.009)	-0.013 (0.011)	-0.013 (0.011)
<i>N</i>	825,115	824,802	825,115	824,802	825,115	824,802	825,115	824,802

[†] Includes controls for year, month, day of week, and prior day's hit rate.

Maximum of homoscedastic and HAC standard errors in parentheses. * indicates $p < .05$, two-tailed.

Table C3: OLS Estimates of discontinuity in hit rate among nonwhite suspects, white suspects, and the difference between the two. All weapon stops 2008-2015.

	Difference in Means	Difference in Means [†]	Linear	Linear [†]	Second Order Polynomial	Second Order Polynomial [†]	Cubic	Cubic [†]
Nonwhite suspects	0.046* (0.002)	0.027* (0.003)	0.026* (0.002)	0.017* (0.003)	0.026* (0.003)	0.017* (0.003)	0.011* (0.003)	0.009* (0.004)
White suspects	0.122* (0.010)	0.102* (0.010)	0.094* (0.014)	0.086* (0.015)	0.068* (0.019)	0.059* (0.019)	0.047 (0.024)	0.045 (0.024)
Diff. in effects	0.076* (0.01)	0.075* (0.01)	0.069* (0.014)	0.069* (0.014)	0.0420* (0.019)	0.0420* (0.019)	0.036 (0.024)	0.036 (0.024)
<i>N</i>	821,532	821,219	821,532	821,219	821,532	821,219	821,532	821,219

[†] Includes controls for year, month, day of week, and prior day's hit rate.

Maximum of homoscedastic and HAC standard errors in parentheses. * indicates $p < .05$, two-tailed.

Figure C1: **Differences in Treatment Effects by Racial Makeup of Block Group:** The figure shows the differences in treatment effects between block groups that are above/below the median % white. Positive estimates indicate that the treatment effects were larger in high-% white block groups.

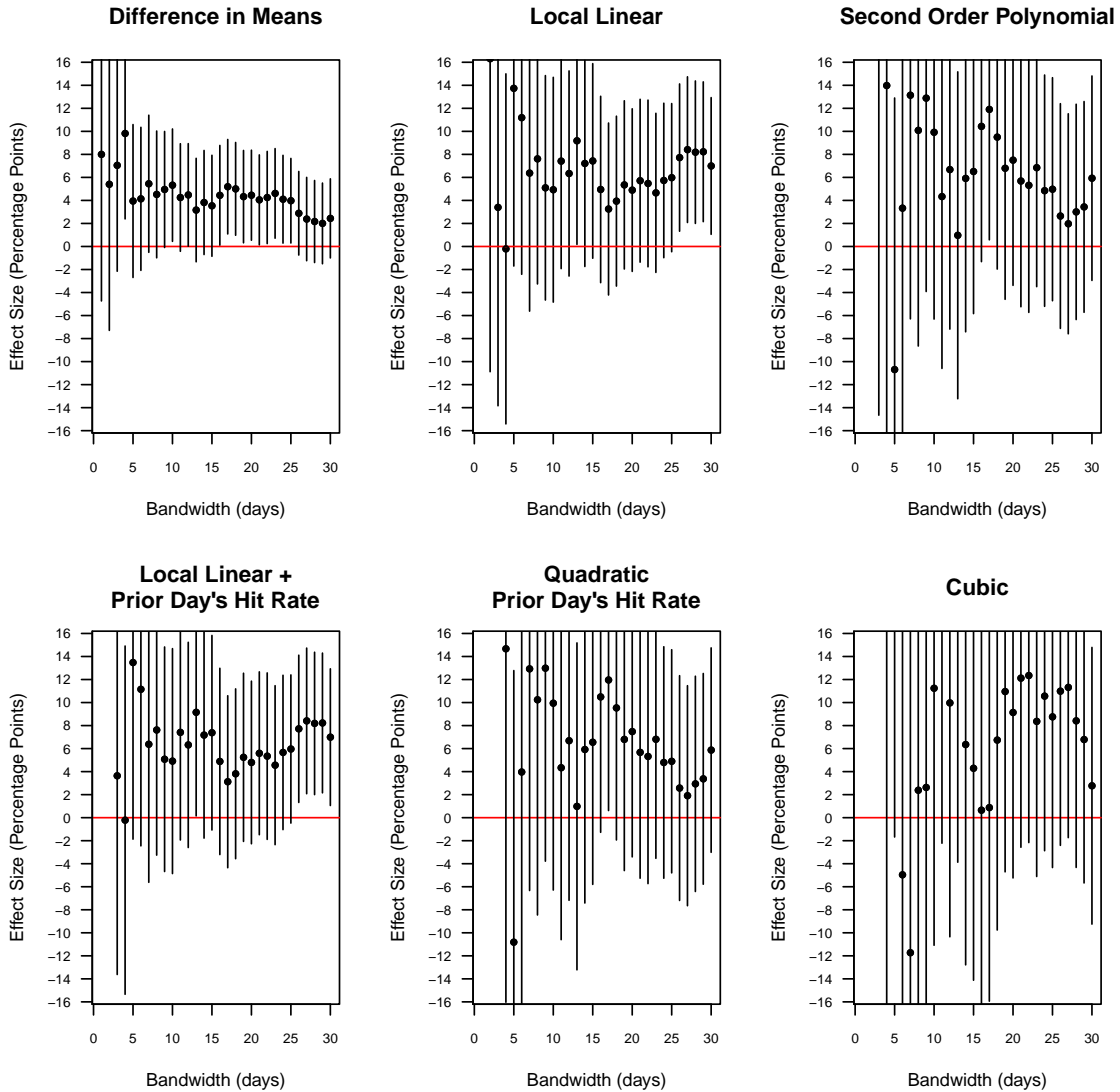


Figure C2: **Differences in Treatment Effects by Precinct Homicide Rate:** The figure shows the differences in treatment effects between low and high-homicide precincts. Positive estimates indicate the treatment effects were larger in high-homicide precincts.

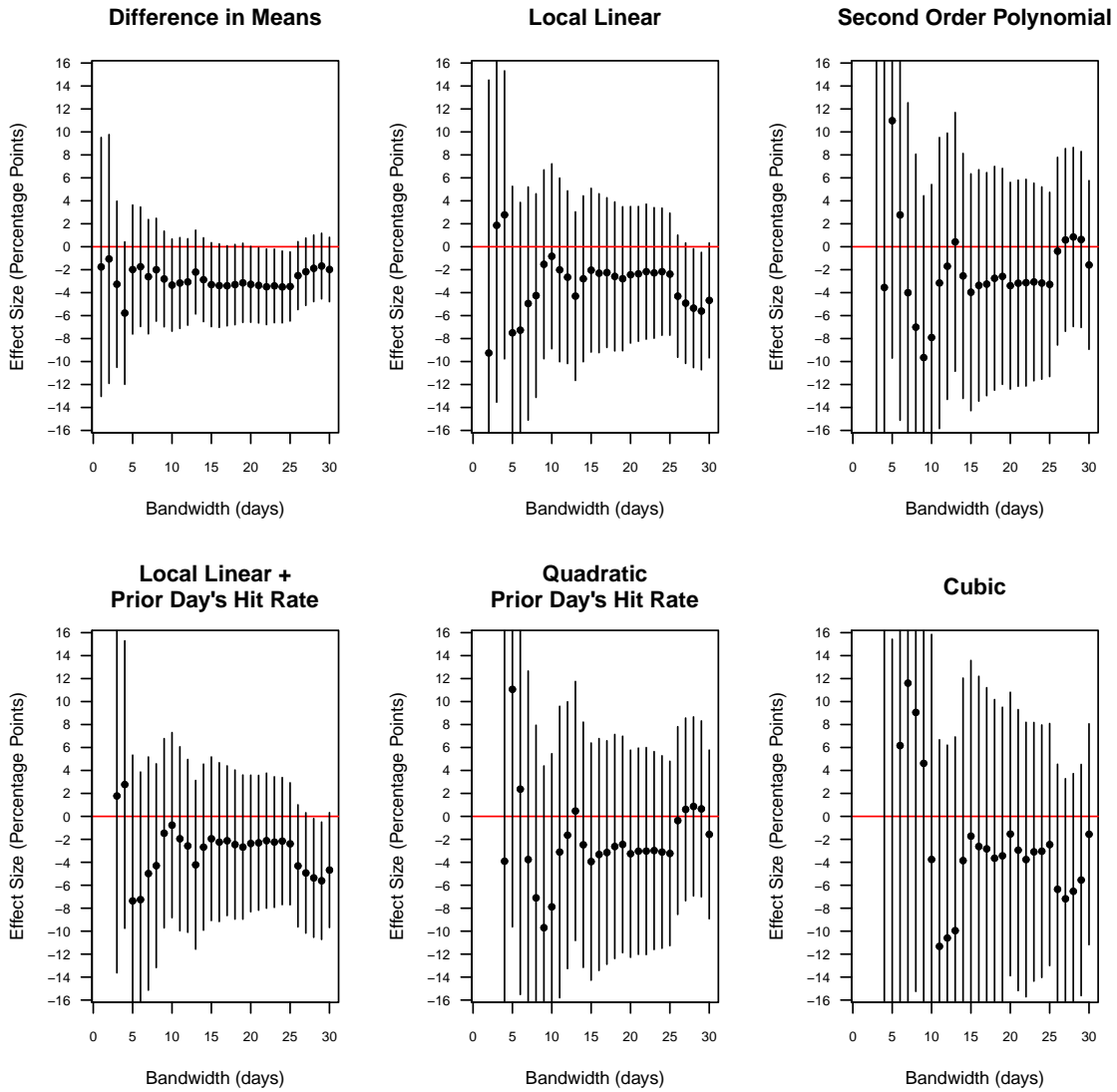
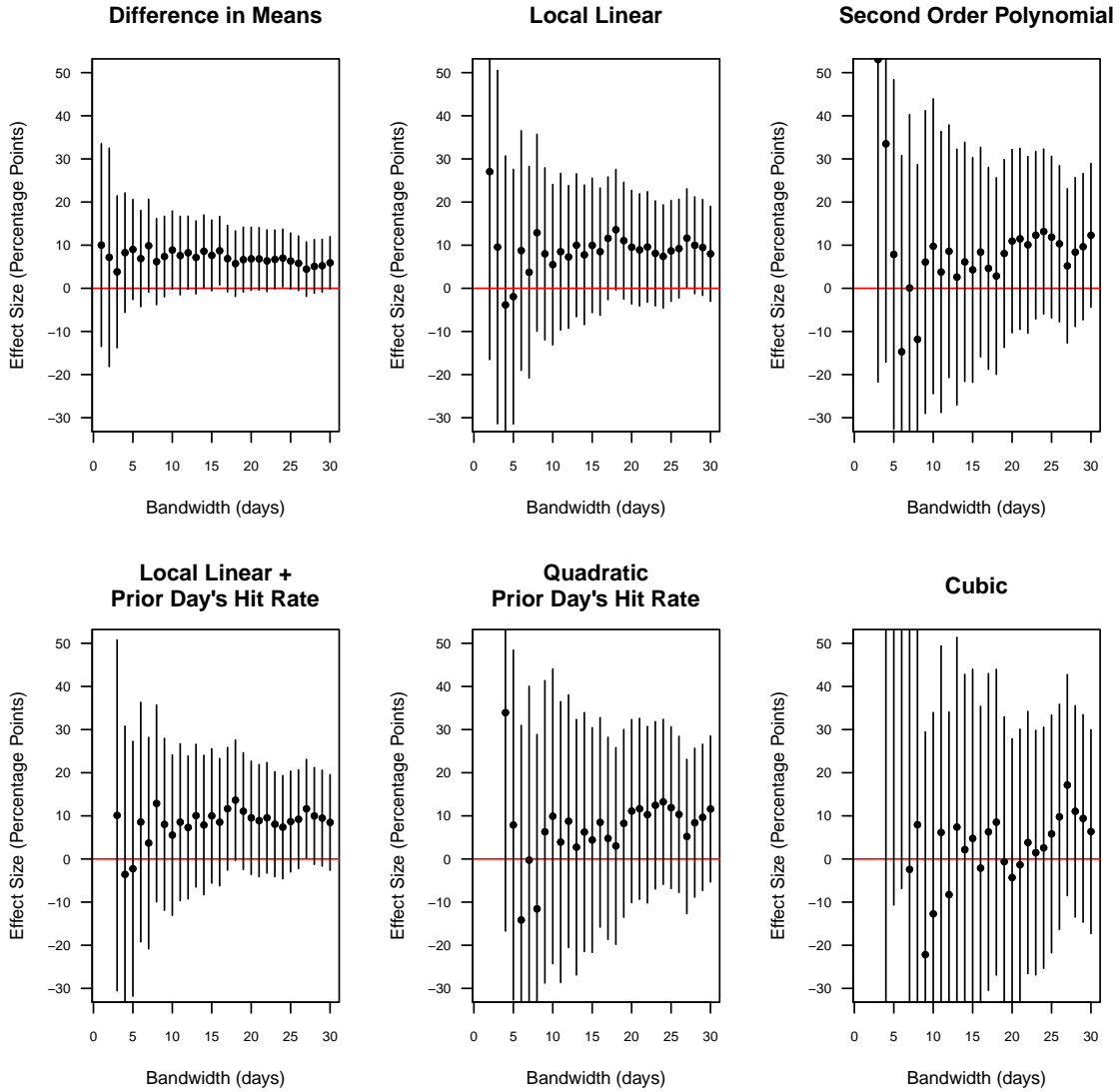


Figure C3: **Differences in Treatment Effects by Race of Suspect:** The figure shows the differences in treatment effects between white and non-white suspects. Positive estimates indicate that the treatment effects were larger among white suspects.



Appendix D: Reporting Bias

Full results for the robustness checks using only data from non-weapon stops, or from weapon stops made by officers in uniform appear in Figures D1 and D2, respectively. Figure D3 shows results of the propensity score analysis cited in the main text using alternative bandwidths. Note: a Kolmogorov-Smirnov test using a 30-day bandwidth rejects the null that the two sets of predicted probabilities generated in the propensity score analysis were sampled from the same distribution ($p < .001$). However, the post-treatment distribution differs only slightly in terms of quantiles, and in the opposite direction than the one implied by the reclassification hypothesis. That is, the predicted probabilities of being labeled a weapon stop (among non-weapon stops) tend to be *smaller* in the post-treatment period than in the pre-treatment period, which is the opposite of the expected result if officers were reclassifying stops that, based on their covariate values, should have been weapon stops. The 25th, 50th and 75th percentiles of the pre and post-treatment distributions are .04, .13, .36, and .036, .11 and .30, respectively. In all, the distributions look highly similar, and do not indicate reclassification.³

³The variables used to predict the probability of being labeled a weapon stop in these logistic regressions were: whether the stop was outside, police precinct indicators, whether the stop was in a public housing/transit/street location (separate indicators), the observation period prior to the stop, whether a suspicious object was seen, whether the suspect fit the description of a known suspect, whether the suspect was seen "casing", whether the suspect was acting as a lookout, whether the suspect was wearing clothing associated with criminal behavior, whether drug activity was witnessed, whether furtive movements were displayed, whether violent activity was witnessed, whether the suspect had a bulge in his clothing, "other", whether the suspect was close to a known offense, whether the suspect was associating with known criminals, whether the suspect changed direction at the sight of the officer, whether the suspect was in a high crime neighborhood, whether the time of day fit the suspected crime, whether sights and sounds of criminal behavior were noticed, a second "other" category, the suspect's sex, race, age, height, weight, hair, build, whether the officer was in uniform, the day of the week and the hour of the day.

Figure D1: The panels below display the immediate change in the weapon recovery rate among stops where a crime other than “criminal possession of a weapon” was suspected using between 1 and 30 days of data on either side of the intervention.

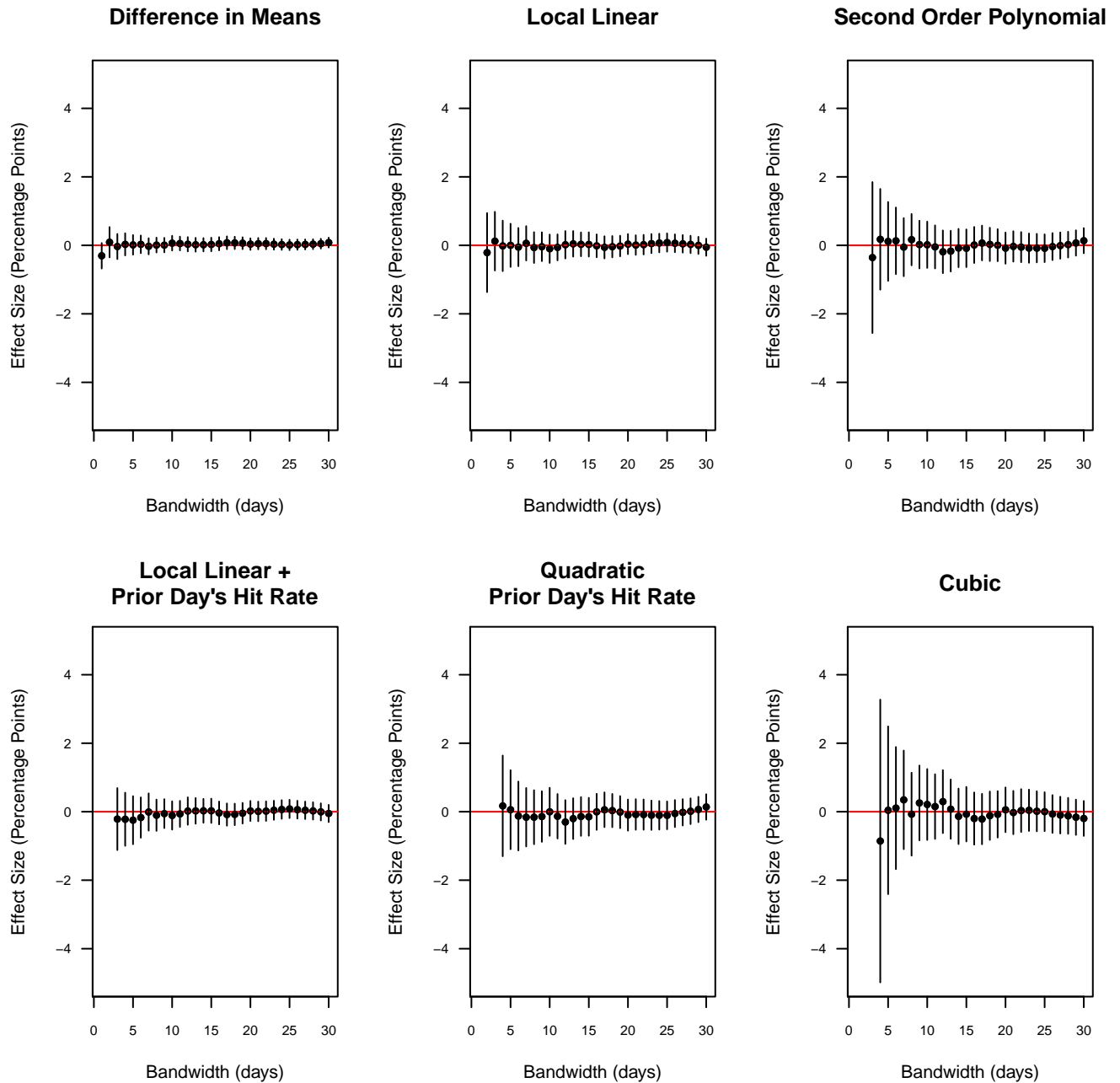


Figure D2: Estimates of the change in the weapon recovery rate at the point of the intervention for stops by uniformed officers only using between 1 and 30 days of data on either side of the intervention.

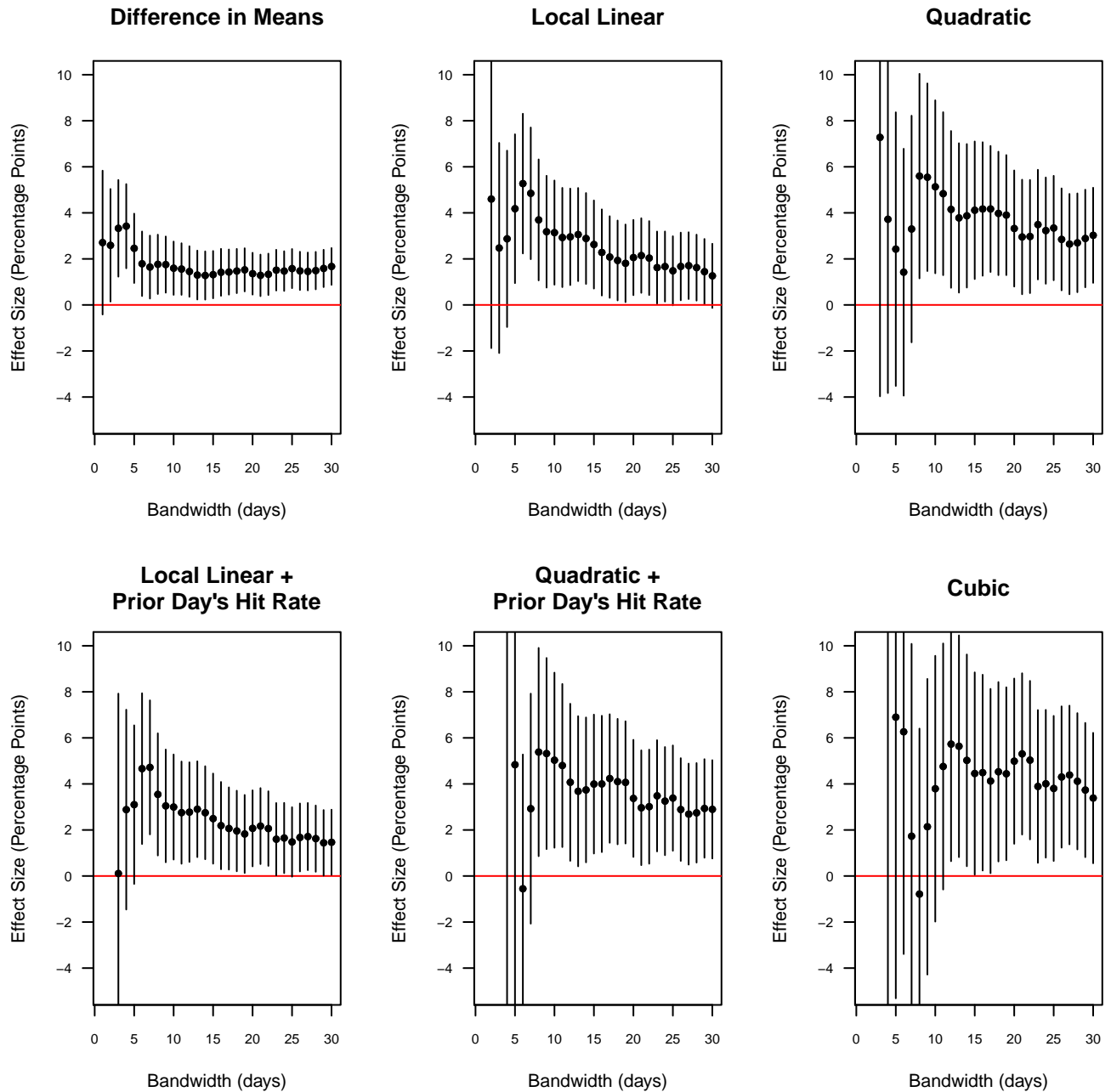


Figure D3: The panels below display the distributions of predicted probabilities (propensity scores) of being labeled a weapon stop using pre-and-post treatment observations at various bandwidths *among observations that were not labeled weapon stops*. A logit model was fit to the pretreatment data to produce the pretreatment distribution, and the parameters it generated were used to estimate predicted probabilities for the post-treatment observations.

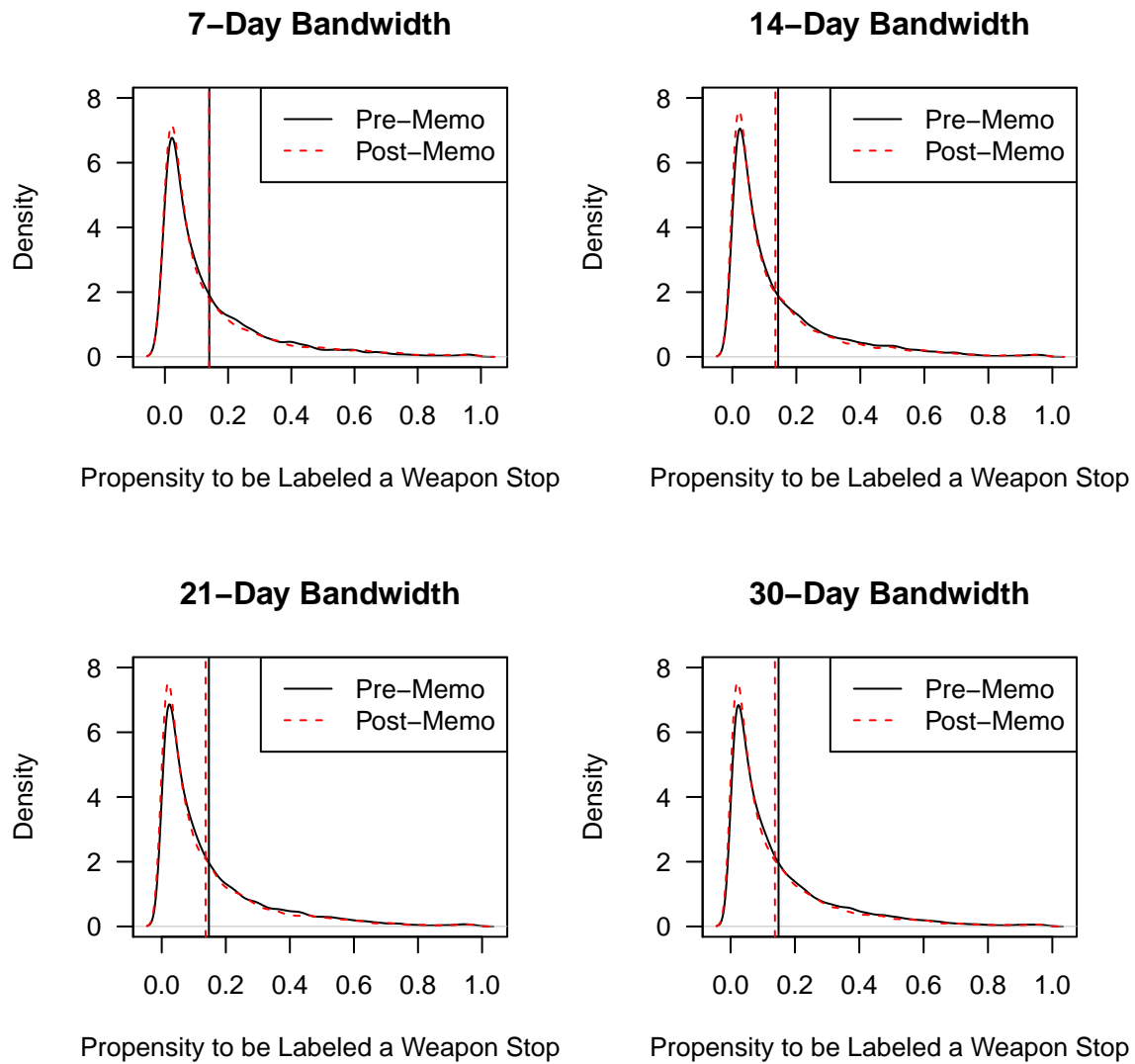


Figure D4: The figure displays the rate at which suspects refused to show identification during a stop over time. According to journalistic accounts (Rayman 2013), officers wishing to report stops that never occurred would often mark this field positively, so it can be used as a rough proxy for the prevalence of this form of data manipulation. If officers suddenly lowered the rate of this practice on the day of the intervention, that could produce an artificially higher hit rate. We see the rate spikes in 2011, the height of SQF in New York, but does not decline at the treatment boundary.

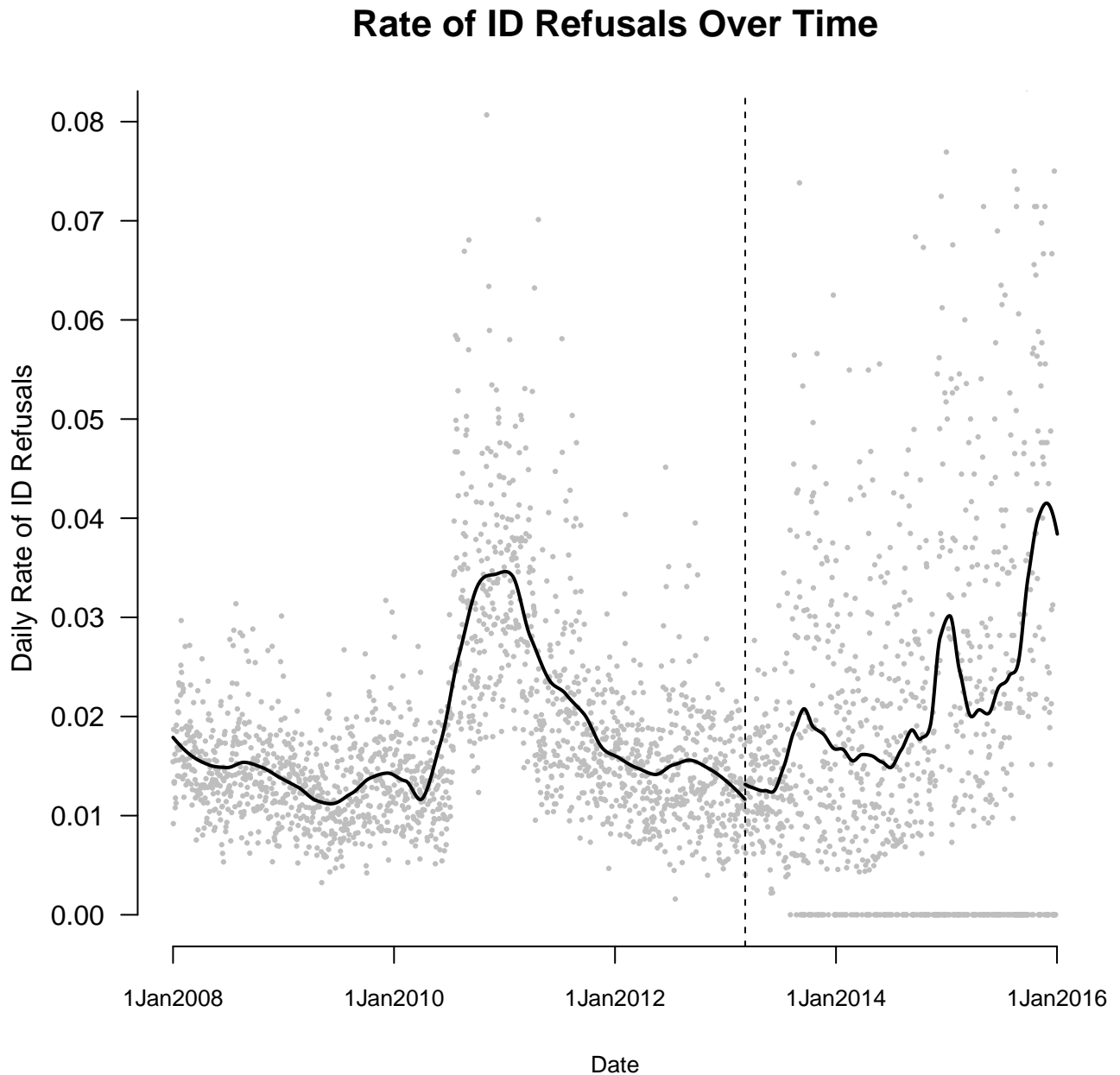
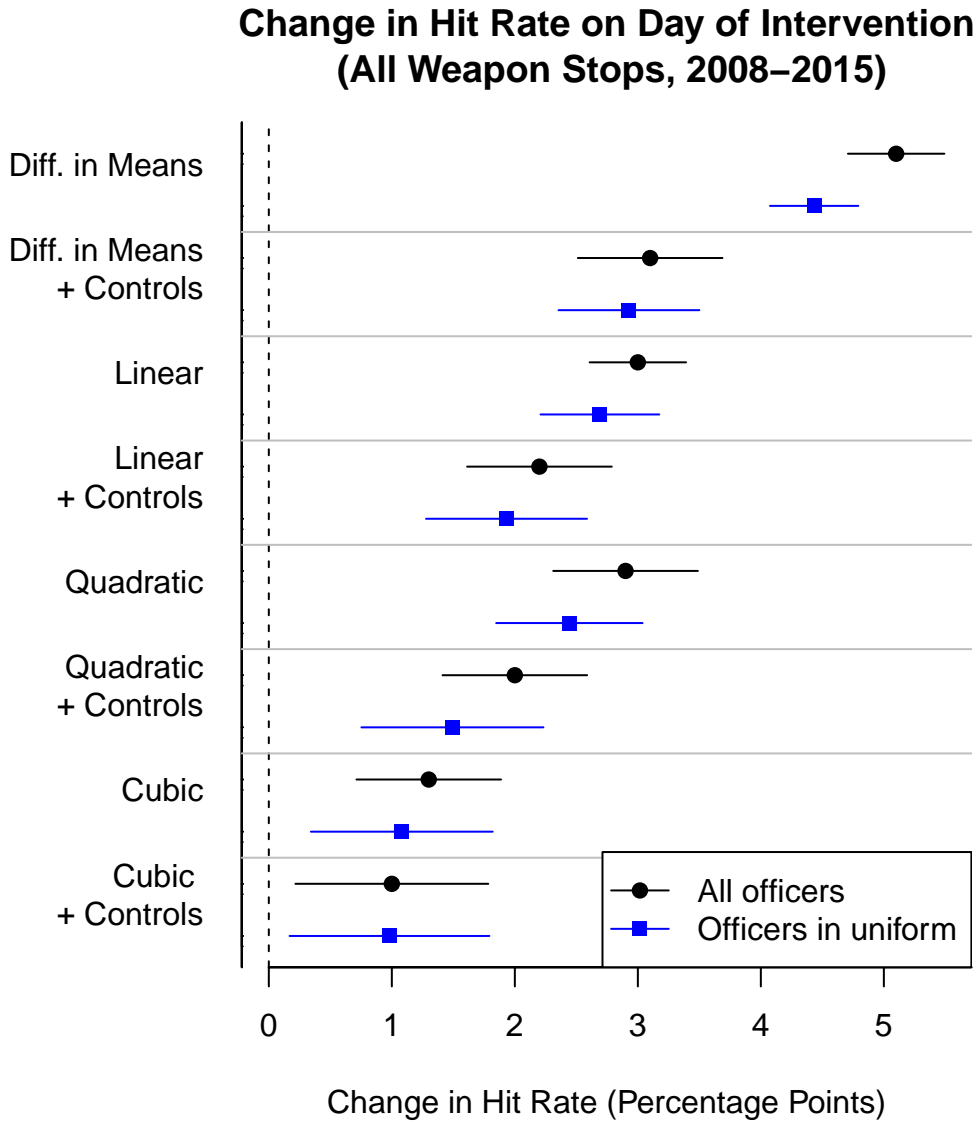


Figure D5: The compares treatment effects using all weapon stops to those estimated using only stops made by officers in uniform—a subset of the data where intentional censoring is unlikely. The treatment effects are highly similar, especially once time trends on either side of the treatment boundary are modeled.



Appendix E: Placebo Checks

Since we know the day the memo was released, concerns over whether treated and untreated observations were coded according to the correct cutoff date are mitigated. However, we might wonder how often using alternate cutoff dates in the data would produce discontinuities similar to the ones observed using the memo release date. Figures E1 and E2 display the distributions of discontinuities computed using every *other* day in the pretreatment data as the hypothetical cutoff date. Figure E1 performs this exercise using 15 and 30-day bandwidths, and Figure E2 does so using all available pre-treatment data on either side of each hypothetical cutoff. The dotted vertical lines denote the middle 95% of these null distributions and the solid vertical lines denote the observed effect using the actual memo release day as the cutoff.

As the results show, these alternative estimates using 15-day bandwidths produce discontinuities comparable to the observed treatment effects generated using the actual intervention date fairly often. However, Figure E2 shows that once the variance in these estimates is reduced by using a larger 30-day bandwidth, the observed treatment effects fall well outside the middle 95% of the estimates in the null distribution, which contains estimates typically at or below 1 percentage point. Using all available data on either side of the hypothetical cutoff dates reveals that the observed treatment effects are highly unusual relative to those in the null distributions, which were again often 1 percentage point or less (see Figure E2). Taken together, these results indicate that the treatment produced an effect much larger than would be generated by chance after randomly picking an alternative intervention date.

Figure E1: **Local placebo check:** Observed effects and null distributions using alternative cutoff dates. Row 1 estimated using a 15-day bandwidth. Row 2 estimated using a 30-day bandwidth. Dotted lines denote the middle 95% of the distribution. Solid red lines denote the observed treatment effects.

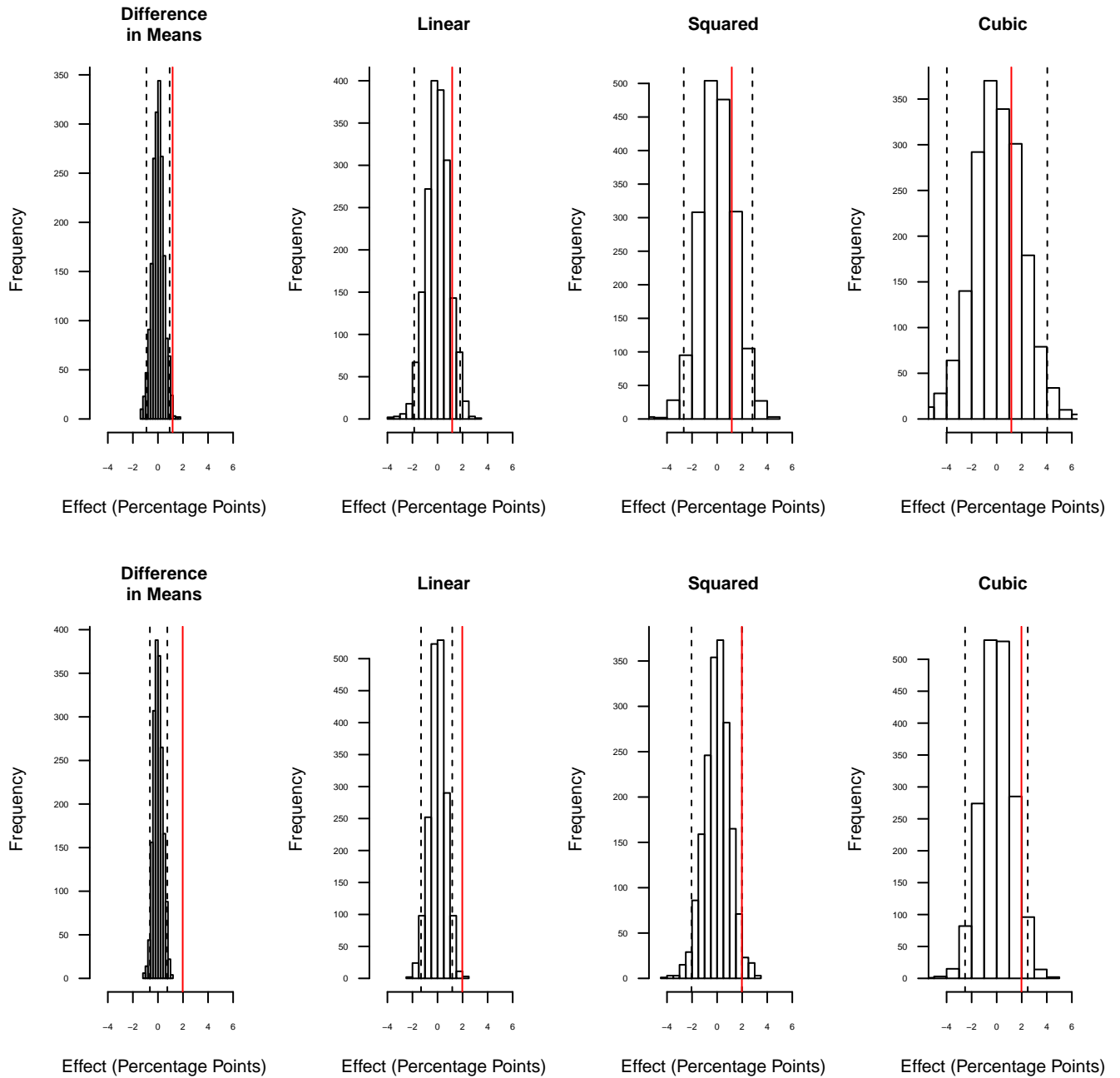
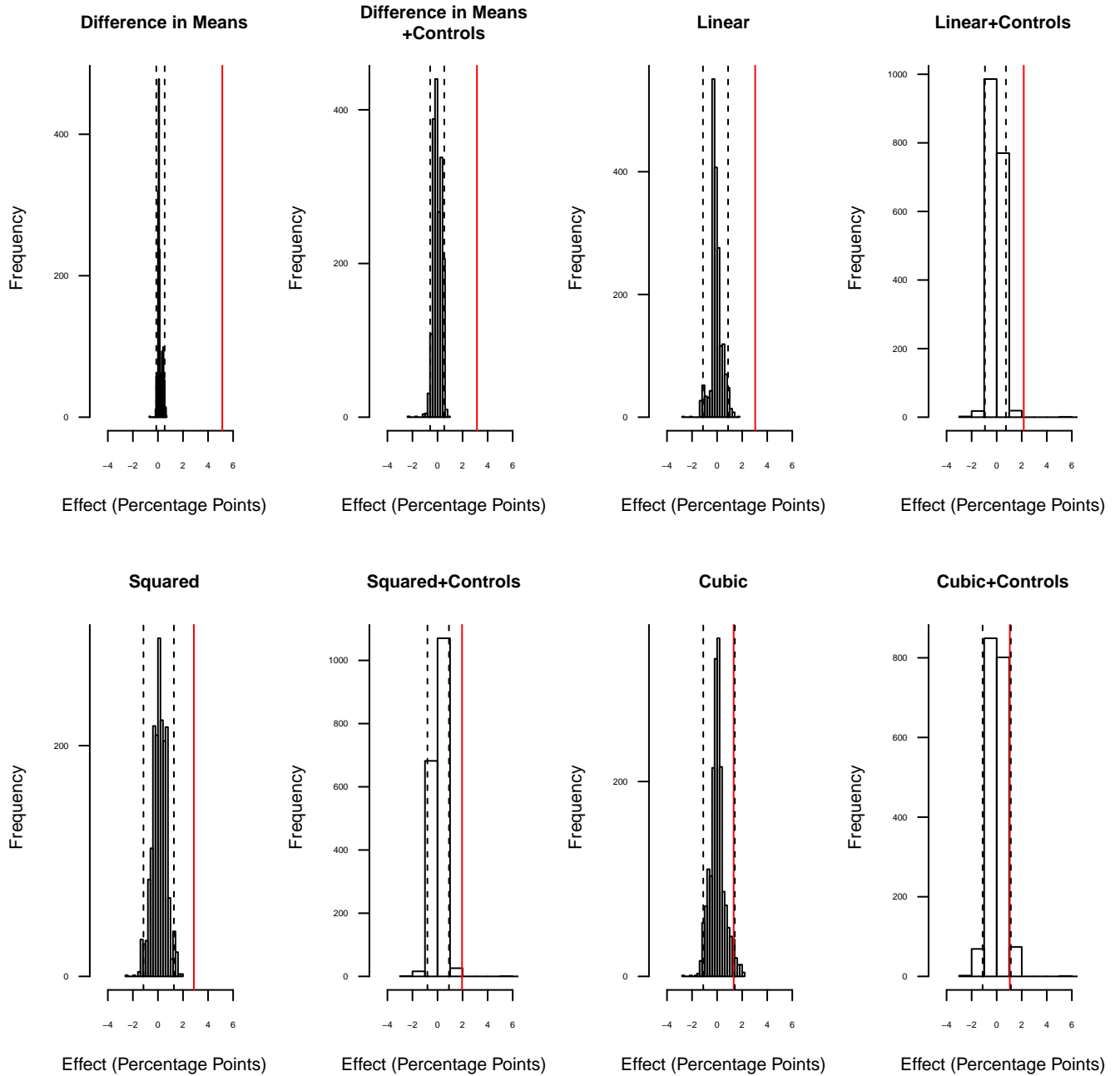


Figure E2: **Global placebo check:** Observed effects and null distributions using alternative cutoff dates and all available data on either side of hypothetical cutoff (prior to actual intervention date). Dotted lines denote the middle 95% of the distribution. Solid red lines denote the observed treatment effects.



Anticipatory Behavior

One potential threat to validity would exist if officers somehow anticipated the new order and changed their behavior before it was given. For officers to anticipate this order in the days and weeks prior to its arrival, it would have to have been planned at least that far in advance. But the memo was likely released in response to a court brief filed just one day earlier by the plaintiffs in the lawsuit related to this policy, as discussed on p. 9 of the manuscript. In addition, given that the treatment appears to have increased the hit rate, any anticipation of the treatment was also likely to raise the hit rate prior to the intervention. If this occurred, it would therefore lead to *underestimates* of treatment effects, meaning the true effects of the intervention are likely even stronger.

To check for signs of anticipatory behavior in the data, Figure E3 displays a loess fit of the hit rate using 200 days of data before and after the intervention. As the figure shows, no such increase is apparent in the pre-treatment period. As a more formal test of this possibility, we can also conduct placebo tests at various bandwidths using the 60 days of data prior to the intervention, with the 30th day prior acting as a hypothetical indicator for treatment.⁴ If officers were anticipating the memo and altering their behavior ahead of time, we should expect to find positive discontinuities using this pre-treatment data and placebo cutoff date. But as Figure E4 shows, there is no evidence of such increases. If anything, it appears the hit rate may have been falling slightly after the placebo intervention date, making the sudden increase on March 5 all the more compelling.

⁴These tests use the same number of days of data and techniques used in the tests of local discontinuities in the main text. Note: this placebo test is recommended in Imbens and Lemieux (2008, 632).

Figure E3: Loess estimator of hit rate fit to daily hit rates ± 200 days from the intervention.

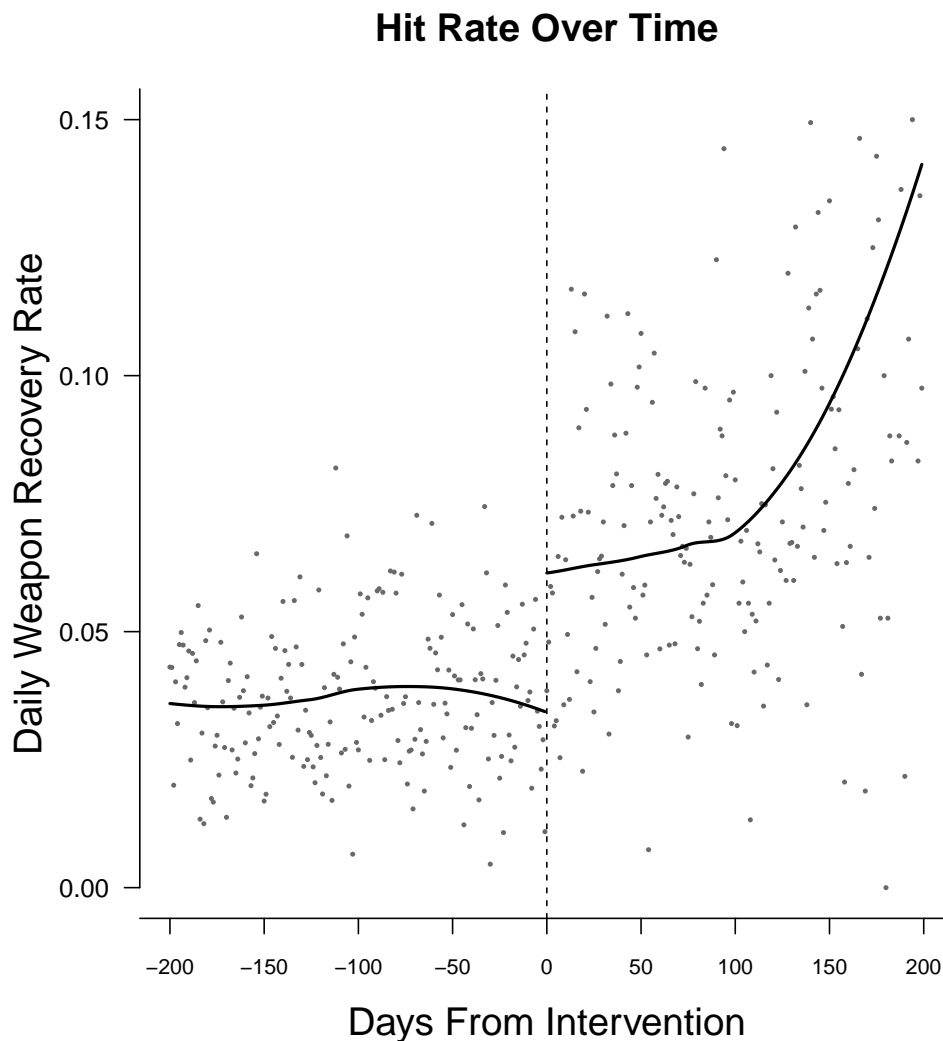
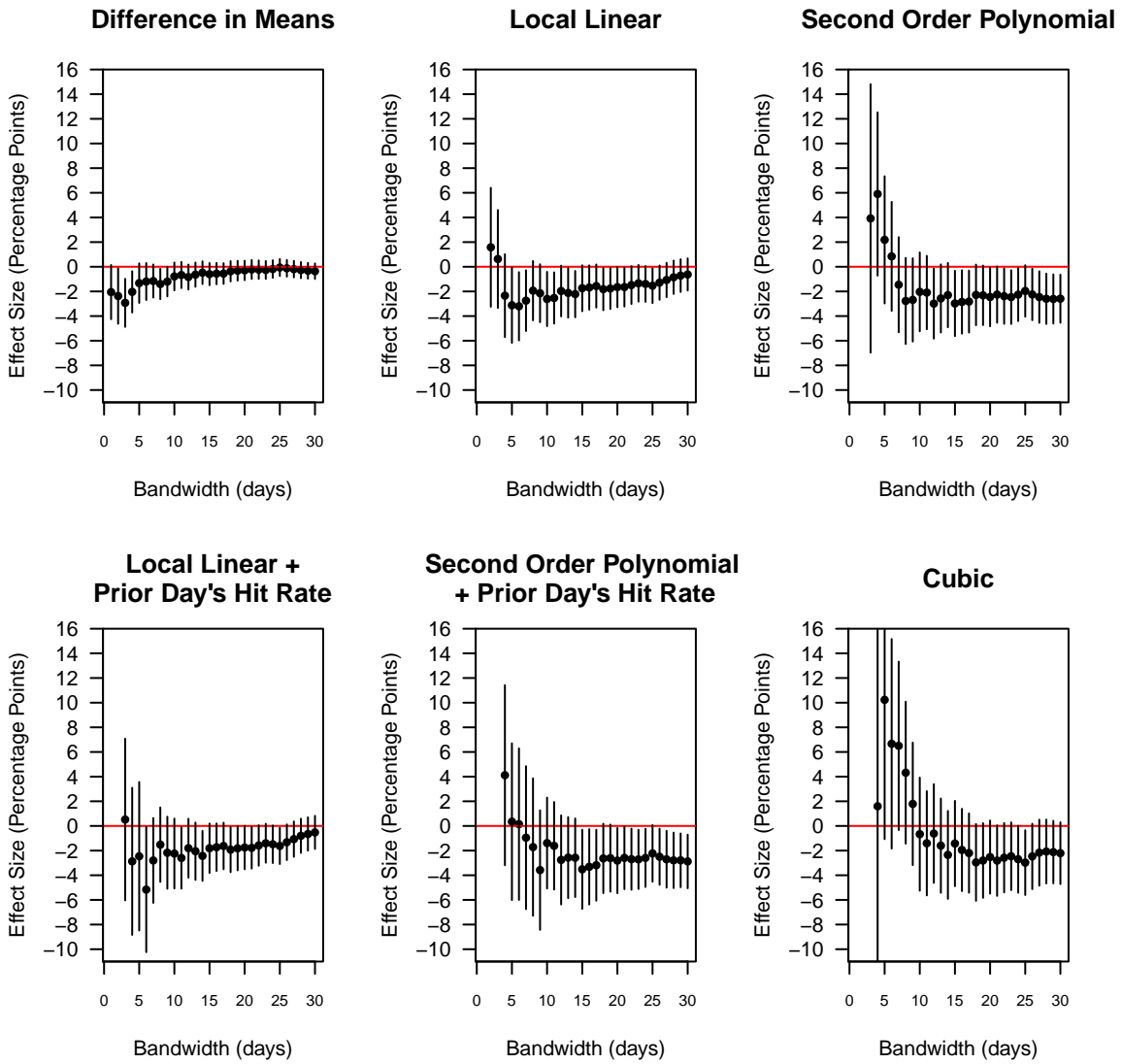


Figure E4: Placebo estimates of the discontinuity in the hit rate using 60 days of data prior to the intervention, and the 30th prior day as the placebo date of treatment.



The Bronx Placebo Test

Placebo tests of whether other orders did or did not produce similar results can be useful for validating causal mechanisms. To construct such a test, one must learn the details of other interventions as well as the exact date of their occurrence. With these constraints in mind, I identified an event suitable for a placebo test, the results of which support the argument that the treatment was effective because it increased the threat of having poor performance scrutinized by *superiors*, who could readily dole out sanctions.

On Jan. 8, 2013, a federal judge ruled that the department's "Clean Halls" initiative, which used SQF to stop "suspicious" looking individuals in apartment complexes primarily in the Bronx, was being conducted improperly (Golding 2013). In her decision, Judge Shira Scheindlin wrote that officers must gather a higher standard of evidence beyond a mere "hunch" based on crime rates in the area or the time of day in order to conduct stops legally. But in a statement, then-NYPD Commissioner Ray Kelly made clear that he did not support the judge's reasoning, saying, "Some may take for granted the safety provided by doormen who routinely challenge visitors to their apartment buildings. . . . The NYPD is fully committed to doing so in a manner that respects the constitutional rights of residents and visitors. Today's decision unnecessarily interferes with the Department's efforts to use all of the crime-fighting tools necessary to keep Clean Halls buildings safe and secure."

Thus, this ruling offers a chance to test whether a similar order to better justify the reasons for making a stop that was *not* supported by NYPD commanders produced similar results. The fact that the judge's order occurred just two months prior to the actual intervention date and not years earlier is also valuable, since the policing environment was likely to be roughly similar in many respects. Table E1 and Figure E5 shows the estimated discontinuities in the hit rate using Jan. 8, 2013 as a placebo intervention day among stops made in the Bronx. As the results show, there is no robust evidence that this ruling produced changes in the hit rate. Without explicit new orders from their direct superiors, who could credibly convey the

threat of sanction for poor performance, officer behavior appears to have remained constant.

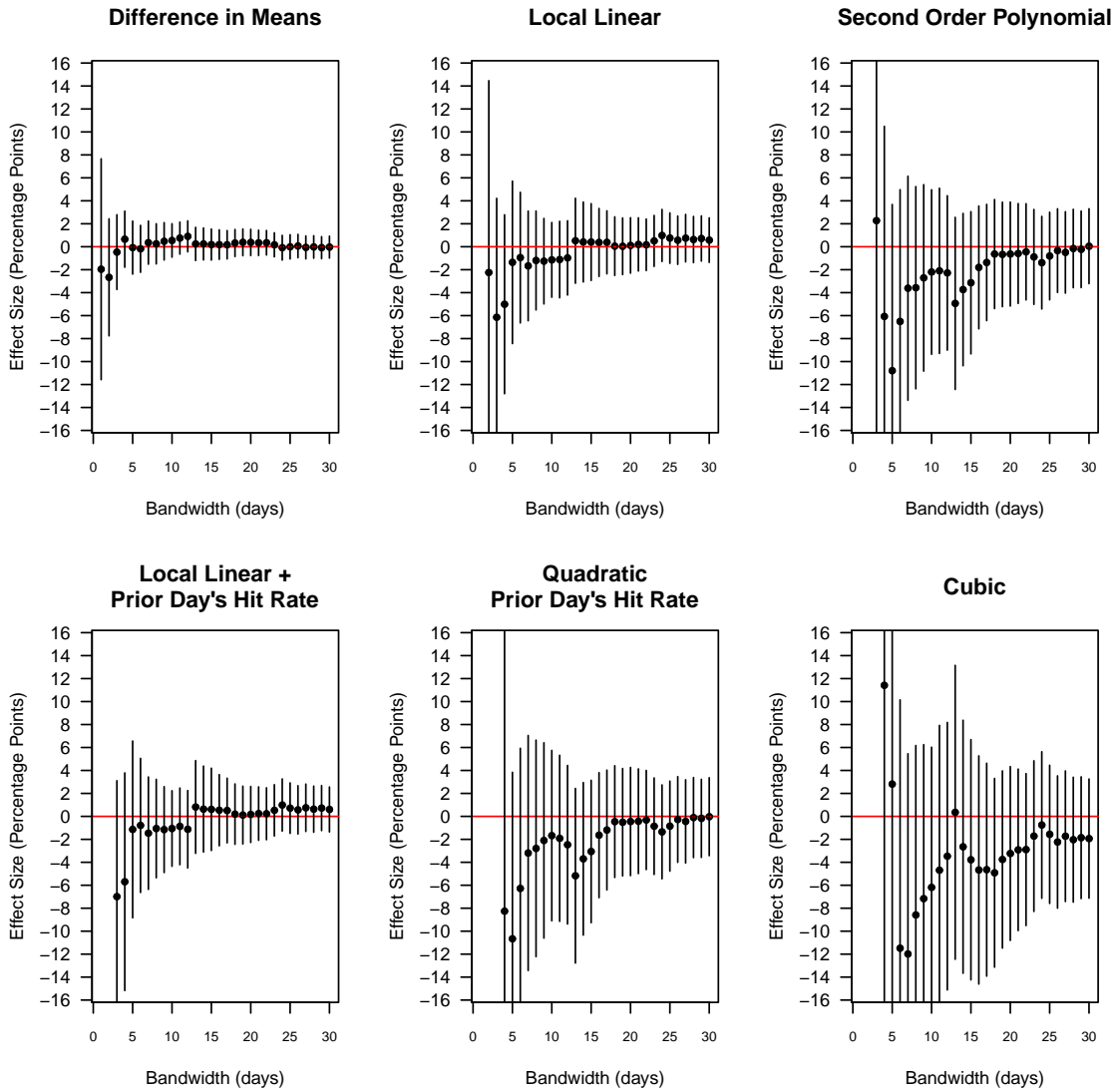
Table E1: OLS Estimates of Discontinuity in Bronx Using Placebo Cutoff Date, All Weapon Stops Prior to March 5, 2013

	Diff. in Means	Diff. in Means [†]	Linear	Linear [†]	Quadratic	Quadratic [†]	Cubic	Cubic [†]
$\hat{\tau}$	0.006* (0.003)	0.000 (0.007)	0.003 (0.005)	0.000 (0.008)	0.002 (0.008)	0 (0.010)	0.013 (0.010)	0.009 (0.012)
N	238,729	238,654	238,729	238,654	238,729	238,654	238,729	238,654

[†] Includes controls for year, month, day of week, and prior day's hit rate in the Bronx.

Maximum of homoscedastic and HAC standard errors in parentheses. * indicates $p < .05$, two-tailed.

Figure E5: The figure shows the estimated discontinuities in the hit rate among stops made in the Bronx using Jan. 8, 2013, the date of the “Clean Hallways” court ruling, as the placebo intervention date.



References

- Calonico, Sebastian, Matias D. Cattaneo, and Rocio Titiunik. 2014. "Robust Nonparametric Confidence Intervals for Regression?Discontinuity Designs." *Econometrica*, 82(6): 2295-2326.
- Efron, Bradley, and Robert J. Tibshirani. 1994. *An Introduction to the Bootstrap*. CRC Press.
- Floyd v. New York City Trial Updates. 2013. "Center for Constitutional Rights." Retrieved November 28, 2015 (<http://ccrjustice.org/floyd-v-new-york-city-trial-updates>).
- Goel, Sharad, Justin M. Rao and Ravi Shroff. 2016. "Precinct or Prejudice? Understanding Racial Disparities in New York City's Stop-And-Frisk Policy." *Annals of Applied Statistics*, 10(1), 365-394.
- Golding, Bruce. 2013. "Judge orders 'immediate cease' to NYPD's stop-and-frisk policy in Bronx 'Clean Halls' building." Jan. 8, *The New York Post*. <http://nypost.com/2013/01/08/judge-orders-immediate-cease-to-nypds-stop-and-frisk-policy-in-bronx-clean-halls-building/>
- Imbens, Guido and Karthik Kalyanaraman. 2011. "Optimal Bandwidth Choice for the Regression Discontinuity Estimator." *The Review of Economic Studies* rdr043: 1-28.
- Imbens, Guido W., and Thomas Lemieux. 2008. "Regression Discontinuity Designs: A Guide to Practice." *Journal of Econometrics* 142(2): 615-635.
- Hainmueller, Jens, Jonathan Mummolo, and Yiqing Xu. 2016. "How Much Should We Trust Estimates from Multiplicative Interaction Models? Simple Tools to Improve Empirical Practice." http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2739221
- Rayman, Graham A. 2013. *The NYPD Tapes: A Shocking Story of Cops, Cover-Ups, and Courage*. London: Macmillan.