Can You Build a Better Cop?  
Experimental Evidence on Supervision, Training, and Policing in the Community

Emily Owens¹  
David Weisburd²  
Geoffrey Alpert³  
Karen L. Amendola⁴

August, 2016⁵

ABSTRACT

Police officers who are actively engaged with the public can reduce crime through general deterrence and by arresting criminals. At the same time, excessive arrests and the use of force by officers can reduce public trust in the integrity of the police. To date, there is scant evidence as to how police departments can successfully train officers to balance these two goals in the field. Drawing on psychology and economics, we developed a low-cost, on-the-job training program intended to encourage officers to “slow down” their thought process during citizen encounters and engage in behaviors consistent with procedural justice. After randomly assigning a select group of officers to training, we find that those who participated in training were as engaged in the community as similarly situated officers, but were less likely to resolve incidents with an arrest and were less likely to be involved in incidents where force was used. These changes were most evident among officers who worked in areas with a modest level of risk.

Keywords: Policing, Job Training, Procedural Justice, Field Experiments

JEL Codes: K42, H11, M53

¹University of California, Irvine, Department of Criminology, Law and Society and Department of Economics, Irvine CA: 92697 egowens@uci.edu  
²George Mason University, Department of Criminology, Law, and Society, Fairfax VA: 22030; Hebrew University, Institute of Criminology, Jerusalem, Israel dweisbur@gmu.edu  
³University of South Carolina, Department of Criminal Justice and Criminal Justice, Columbia SC: 29208; Griffith University, School of Criminology and Criminal Justice, Brisbane, Australia geoffa@mailbox.sc.edu.  
⁴Police Foundation, Washington DC: 20036 kamendola@policefoundation.org  
⁵We thank Maria Valdovinos and Earl Hamilton for assistance at all stages of this project, along with Scott Decker, Matthew Freedman, Ben Hansen, John MacDonald, Travis Taniguchi and seminar participants at the University of Pennsylvania-Wharton and the Royal Economic Society Annual Meeting for helpful comments. This research was made possible by the cooperation and enthusiasm of Frm. Deputy Chief Clark Kimerer, Captain James Dermody, Kim Kubie, Mimi Walsh, Peggy Garcia, and Jan Hoyt of the Seattle Police Department and Claudia Gross-Shader of the City of Seattle. We also acknowledge the support and ongoing assistance of our grant monitor, Dr. Brett Chapman whose commitment to this project was unwavering. This project was supported by Award No. 2012-IJ-CX-0009 awarded by the National Institute of Justice. Part of this research was made possible through the use of Cornell University’s Social Science Gateway, which is funded through NSF Grant 0922005. The opinions, findings, and conclusions or recommendations expressed in this publication/program/exhibition are those of the authors and do not necessarily reflect those of the Department of Justice or any other agency. All errors are our own.
1. Introduction

How do law enforcement agencies convince the public that they are benevolent and trustworthy, while at the same time making arrests and using physical force? Officers must strike a delicate balance between issuing warnings and citations, and detaining those accused of violating the law, while at the same time making sure that all citizens feel like they are treated fairly and equitably and are protected from harm. Police researchers have characterized this as an “impossible mandate” for departments that requires being simultaneously perceived as “warriors,” who aggressively attack criminals, and as “guardians,” who nurture and protect residents (Manning 1977).

High arrest rates and publicized incidents of force used against citizens by police, especially when interpreted by the public as excessive, and when crime rates are low and the social costs of having a criminal record are high, can make it difficult for police departments to find the optimal combination of their warrior and guardian roles (Rahr and Rice 2015, Stoughton 2015). Indeed, dissatisfaction with police practices led to a series of nationwide, occasionally violent, protests in the summers of 2014, 2015, and 2016. The immediate catalysts for these protests were specific incidents in which officers caused the death of a citizen (and were not perceived to have legitimately done so), but the sustained nature of the public outcry is generally understood to be a response to a perceived militarization of police in the US; an overemphasis on the warrior role, where officers stop, detain, or use force against individuals, particularly racial and ethnic minorities, without adequate cause or explanation.¹

Academics in law, psychology, and criminology argue that part of the solution to the tension between these two roles is to use “procedural justice,” treating citizens fairly and with dignity, no matter what the outcome (Meares et al 2016). As defined in Tyler (1988),

¹ For example, see Friedersdorf, Conor, “The Case for Police Reform is Much Bigger than Michael Brown,” The Atlantic (November 26, 2014).
procedural justice is based on seven components: the officer’s perceived motivation, honesty, and ethicality, opportunities for representation of all parties, opportunities for error correction, the quality of the officer’s decision, and the officer's bias in reaching a decision. In theory, police departments that implement procedural justice practices will build social capital and be trusted by the communities they serve. This will then lead to increased witness and victim cooperation, and ultimately lower crime rates. While this idea has been heavily promoted at the federal, state, and local levels, there is little evidence-based guidance for police chiefs to train their officers to behave in a way that is consistent with these goals, and thus far only high cost, intensive interventions have shown any sign of changing actual officer behavior in the field.

In cooperation with the city of Seattle and the Seattle Police Department, we developed a low cost, low intensity supervisory program inspired by the principles of procedural justice. Building on recent advancements in psychology and economics, the program was aimed at affecting the way officers gather and process information on the job. The aim of the program was to influence the way that officers think about even the most mundane aspects of their job, with the goal of reducing the frequency with which officers engaged in behavior that could be potentially viewed by the public as unjust. It involved short supervisory meetings with officers who were identified to be at higher risk of negative encounters with the public.

Our training program differs from standard police supervision and training programs in three ways. First, during these meetings officers and their supervisors discussed a relatively benign event which would typically not be subject to discussion under standard practice. Second, rather than focusing on outcomes, officers were prompted to reflect on their thought processes and actions during these encounters. Finally, supervisors were trained to model

---

3 See, for example the 2015 report of the Major Cities Chiefs Association, Major County Sheriffs Association, and FBI at https://majorcitieschiefs.com/pdf/news/community_engagement_5_27_15.pdf
some of the central components of procedural justice during these meetings. Over a six-month period, officers were randomly enrolled in the program, which the officers and their supervisors generally reported as non-invasive.

We evaluated the impact of this program using direct measurements of officer performance in the field. It is important to reiterate that we did not directly observe whether or not officers used procedural justice, but rather the more traditional, observable, and quantifiable measures of officer “output” that their line supervisors, the Federal Government, and the public care about.4 We found that, in the six weeks following a supervisory meeting, treated officers were as active in the community as untreated officers along multiple dimensions, but were 12% less likely to resolve incidents with an arrest and were almost 50% less likely to be involved in use of force incidents. We found that the largest reduction in arrests occurred among officers who worked beats where there was a moderate level of “predicted risk,” which we defined using the frequency with which other officers used force, were injured, or were the subject of citizen complaints after working in that area. While we are limited in our ability to identify precise mechanisms, the timing of these reductions, along with officer feedback on the meetings, suggest an important cause of this behavioral change was skill development that occurred during the meeting rather than Hawthorne effects or concerns about being monitored more generally. Overall, we conclude that these non-disciplinary, proactive, supervisory meetings are a promising strategy for improving community-police relations.

The paper proceeds as follows: In section 2, we provide background on police training and procedural justice, along with research in labor economics and psychology that informed our training protocol. Section 3 describes how we identified officers who would be eligible for the experiment, and we describe our training “engagement” in section 4. We describe the

---

4 Ibid.
performance measures we used to evaluate our intervention and our simple analytic model in sections 5 and 6, respectively. Our experimental findings are presented in section 7, and section 8 concludes with a short discussion.


The federal government has, for some time, defined “good” policing as policing with procedural justice (e.g. Reno 2001). In economic terms, the procedural justice model of policing is essentially one about the credibility of officers as agents of a benevolent utilitarian social planner; citizens must perceive that (1) they are being treated fairly relative to others (motivation, honesty, and ethicality), (2) they are given the chance to explain or defend their behavior (opportunities for representation and error correction), and (3) their explanation is taken into account before any police action is taken (perceived officer bias and decision quality) (Tyler 1988, Sunshine and Tyler 2003). While procedural justice is technically neutral on the frequency with which officers make arrests or use force, in practice the two are inextricably linked (Kuhns et al 2011). Arrests and use of force are consistent with procedural justice if, and only if, they are objectively proportionate to the specific circumstances of the specific citizen encounter in which the officer uses those actions. When the actions appear to be arbitrary, racially disproportionate, and/or unjustified by the actual situation, they are more likely to be viewed by the public (and the Department of Justice) as excessive.

Existing empirical assessments of the relationship between procedural justice and attitudes toward police have been based on cross-sectional survey research. Essentially all find that higher levels of perceived police quality or legitimacy are associated with higher rates of citizen compliance, cooperation, and law abidingness, feelings about procedural justice are strongly associated with attitudes towards officers, and that citizens’ personal experiences of police processes are highly correlated with their general assessments of the

---

5 There is a separate line of research which focuses on the impact of procedural justice training on the attitudes of police towards procedural justice as a concept (Skogan et al. 2015, Wheeler et al. 2013).
fairness and effectiveness of the police force (Sunshine and Tyler 2003, Tyler 1990, 2001, and 2003, Tyler and Fagan 2008, Tyler and Huo 2002, Tyler and Wakslak 2004, Engel 2005, Nix et al. 2015, Reisig et al. 2007). Evidence from the field is more qualified; an experimental expansion of local community policing offices in Trinidad and Tobago found that residents’ perceptions of safety and the ability of police to prevent crime improved, but the same residents were not more likely to report that police used force in a fair and legitimate way (Kuhns et al. 2011).

Many police academies incorporate training in procedural justice for newly hired officer candidates as a means of promoting the integrity of the department and ensuring the officers act in ways that promote positive community relationships, and the major accrediting body, the Commission on Accreditation for Law Enforcement Agencies, explicitly incorporates procedural justice training as part of its protocol (Alpert et al. 2004, Hass et al. 2015). However, in spite of the widespread belief in procedural justice as a way to improve policing, scant evidence from the field exists to support the idea that academy training in procedural justice or practical on-the-job supervision can encourage officers to change the way in which they interact with citizens. In fact, a 2004 National Research Council report noted that the real-world impacts of police training and education programs in general were essentially unevaluated by social scientists (Skogan and Frydl 2004).6

Two recent studies highlight the practical issues associated with job training programs aimed at changing officer behavior. Rosenbaum and Lawrence (2011) found only modest effects of 20 hours of academy training (of about 1,000 hours) in procedural justice on the behavior of recruits in a role-playing situation. Treated recruits were more likely to say that they intended to use procedural justice concepts in hypothetical situations, but the actual field

---

6 Thanks are due to Wesley Skogan for bringing this to our attention. There is comparatively little attention paid to evaluating the impact of policy changes on the legality or legitimacy of police behavior. Heaton (2011) is a notable exception, which uses a difference in differences strategy to identify a reduction in the rate at which black people are arrested for car theft in New Jersey after a racial profiling scandal lead to increased oversight of traffic stops.
effects are unknown. Mazerolle et al. (2014) found evidence that the use of procedural justice improved perceptions of police legitimacy in the field, but their intervention – having police officers conducting one specific type of traffic stop (sobriety check points) read from a pre-specified script under supervision of a member of the research team and command staff – has limited applicability to the broad tasks of the police, which include various and unscripted types of interactions with the public. The wide variety of situations that officers find themselves in means that, from an organizational perspective, officers have a larger amount of discretion than typical lower-level employees in hierarchical firms (Wilson 1969). This discretion may mute the impact of any training intervention, and potentially limits the external validity of lab studies (where the officer’s actions are entirely observed) with respect to field outcomes (where the officers are essentially unmonitored).7

There are three important problems with the academic model of procedural justice in the practical promotion of “good” policing. The procedural justice focus on the way in which an officer makes a decision, rather than the outcome of that decision, is at odds with traditional policing philosophies and performance measures that are based on order maintenance (Gau and Brunson 2010). Through the use of crime and arrest-focused supervisory tools like CompStat,8 as well as congressional and media attention to the Uniform Crime Reports, department and officer quality is annually evaluated based on crimes, arrests, and clearance rates. Procedural justice and perceptions of officer legitimacy are arguably more nuanced, qualitative concepts, and in the absence of a familiar metric with which to evaluate an officer’s performance, it is difficult for supervisors to promote any sort of behavioral change.

Further, police departments themselves do not typically use procedural justice ideas within their agency (Melekian 2012). Police departments are organized in a hierarchical way,

---

7 More recent research on police supervision supports this concern. For example, Engle (2000) finds that having a supervisor literally present on scene may increase the probability that an officer makes an arrest, but less evidence that more general supervision methods otherwise affects incident outcomes.

8 For more background detail on CompStat, see DeLorenzi et al. (2006)
with individuals either giving or receiving orders depending on their position in the organization. Officers do not expect to have their sergeant’s orders carefully explained or justified, just as sergeants are not in a position to question the fairness of the actions of their lieutenant. In some sense, the promotion of procedural justice techniques requires that officers interact with citizens in a way that does not reflect the way their supervisors typically interact with them (Rahr and Rice 2015).

Finally, research and theoretical insights from economics and psychology suggest that implementing a procedural justice training program that actually changes officer behavior is a complex task. The idea that officers take in and actively use information and opinions from citizens during every encounter, which is the central part of procedural justice, is counter-balanced by canonical ideas in labor and personnel economics about job experience and work productivity.

A central concept in labor economics is that the more a worker preforms a specific job, works at a particular firm, or engages in a particular task, the easier and less time-consuming the actions involved become. This makes an experienced worker more efficient and productive than a similar worker with less experience. The idea that people with more experience develop skills that make them more productive is frequently attributed to Mincer (1958), formally developed in Becker (1993), and recently reviewed in Autor and Handel (2013).

Research in psychology has identified a specific mechanism linking work experience and productivity: diminished cognitive load associated with performing familiar, as opposed to unfamiliar tasks (McDaniel et al. 1988, Rowe 1988). A physical change in the cognitive processing of familiar situations has been established in laboratory studies of how people neurologically respond to new tasks (van Merrienboer and Ayers 2005); when an individual performs a task multiple times, the actual location of brain activity required for that task
shifts, reducing the required amount of cognitive energy. The practical implication of this shift in brain activity is that a person can “automate” their behavior, paying increasingly less active attention to small details and taking small cognitive “shortcuts” to reach the same result (Augoustinos and Innes 2011, Fiske and Taylor 1991, Tversky and Kahneman 1974). In almost all jobs, this automation of routine activities leads to pure efficiency gains - people are simply more efficient at their jobs as they no longer have to think through every step of every task they must perform (Bruya 2010, Ehrenstein et al. 1997, Helie et al. 2010, Yamaguchi and Proctor 2011).

For the special case of police work, automation will lead to better policing in the sense that experienced offers will be, on average, better at anticipating how citizens will behave, but it also works against a citizen’s perception that they are being treated fairly in the event of an arrest. The logic of this insight is straightforward. Officers quickly gain experience in certain types of citizen encounters- loitering, jaywalking, and panhandling are all common disorder offenses that most officers encounter routinely. Officers who regularly patrol particular beats can immediately identify individuals who appear to be intoxicated or in mental distress, or are otherwise in situations that are on the verge of escalating into a criminal disturbance (Chan et al. 2003, Willis 2013). This process has traditionally been referred to in the policing literature as developing “perceptual shorthand” to identify “symbolic assailants” (Skolnick 1966). When more seasoned officers respond to these types of “mundane” events, the very fact that they are experienced means that they will, to some extent, be automating their behavior. Once automation begins, by definition officers are no longer consciously processing all new information provided to them by involved citizens, which is a necessary component of procedural justice in policing.

9 Note that this implies that there is an “optimal” point in an officer’s career, when they should have the most positive impacts on society.
10 This shift in cognitive processing is highlighted by a phrase used by a Seattle officer approaching an apparently transient set of teenagers in a commercial space: “I’ve seen this movie before.”
It is important to note that experienced officers can usually resolve low level situations quickly and without further criminal justice involvement for the citizen. A positive relationship between experience and automation does not mean that, on net, seasoned officers are less likely to be perceived as “fair” than rookies – rather, *ceteris paribus*, deliberately gathering information in response to any incident will lead officers to behave in ways that are more in line with the procedural justice ideal.\(^\text{11}\) Of course, true *ceteris paribus* situations are rare; the inexperienced officer may be less adroit at de-escalation techniques, may be more anxious to use the tactical training obtained in the academy, and may be more eager to “prove themselves” by making arrests, all of which are likely to offset the potential gain of gathering more information.

These basic insights about job performance and cognitive processing have profound implications for understanding the way in which police interact with citizens, and why otherwise competent officers can end up in situations where their actions appear to be disproportionate to the initial circumstances of a citizen encounter.\(^\text{12}\) As soon as officers become so good at their job that they no longer need to actively cognitively process each specific detail of the event, they may also stop actively reflecting on their choice of words with a jaywalker or paying careful attention to the non-verbal communication of a panhandler. In this sense, what would otherwise be called “efficient” job performance is contrary to the goals of procedural justice and perceived fairness in law enforcement.

From a policy perspective, in the rare case where a mundane encounter escalates into something that is more serious, officers who continue to consciously incorporate call

\(^{11}\) A similar logic is presented in Heller et al (2015), which evaluates a cognitive training program aimed at reducing youth violence; for certain young men, an automatic aggressive response to potential threats actually keeps them safe in their neighborhoods. However, encouraging them to slow down their thinking in mundane situations (e.g. why not ask the person for something you want, rather than try to force them to give it to you?) appears to reduce the frequency with which they engage in violent behavior on the street.

\(^{12}\) Examples of this include the 2009 arrest of Dr. Henry Louis Gates, Jr. by Sgt. James Crowley, the 2014 arrest of Dr. Ersula Ore by Ofcr. Stewart Ferrin, the 2014 detention of Wesley Lowery in Ferguson, MO, and multiple use of force incidents occurring in MO, NY, and CA during the summer and fall of 2014.
available information (e.g., the body language and tone of the involved citizens) may actually be able to recognize and diffuse tense situations sooner (Alpert et al. 2004). In the field, this would lead to fewer decisions in which an arrest was necessary, and less frequent need for the officers to use force to regain control of an initially mundane situation that, without their fully noticing, had become more serious. This idea will motivate our analysis, and is a central hypothesis of interest.

The tension between the “fast thinking” associated with learned patterns of behavior and “slow thinking” required for managing citizen encounters in ways that encourage procedural justice is central to understanding how to successfully implement training programs that increase police legitimacy. Our experiment goes beyond prior work in this area, by explicitly recognizing this tension; our supervisory program aimed to encourage even experienced officers to “think slower” and more deliberately gather information during citizen encounters. To be clear, there will always be situations where officers must use force or make arrests. However, to the extent that (otherwise efficient) automation can occasionally lead to the escalation of an initially mundane encounter into one that is more serious, officers who continue to gather and process all information available to them on scene will be more likely to diffuse potentially explosive situations before they occur. Further, to the extent that citizens feel that police are responding to them personally, they may be less likely to respond in a way that escalates the stakes in the encounter. Our experiment provides evidence on whether this approach is actually effective in the real world.

3. Identifying the Experimental Sample Pool

We evaluated the impact of an adapted procedural justice training program in Seattle, Washington. During our study, the Seattle Police Department was undergoing dramatic change; the department was under the oversight of federal monitor, a new mayor who ran on a police reform platform came into office, and there was high turnover among the police
leadership. By employing an experimental method where we randomly assigned a subset of officers to our intervention, we were best able to control for the potential influences of these factors, as both treated and control officers were subject to the same fluctuations in political and criminal climate. Note that in order to balance these external influences across treatment and control, it is necessary to observe treated and control officers at the same point in time.

In addition, not all officers engage in, or are at risk of engaging in, the sort of citizen encounters where additional procedural justice training and cognitive engagement might make a difference in outcomes. For example, officers who work in very low crime residential neighborhoods may issue very few citations, let alone make arrests or use force. Ideally, we would evaluate the impact of the training program in a set of officers who, in the future, may be more likely to be involved in a serious encounter, but had not necessarily been involved in one yet. Intuitively, the pool of officers whom we wanted to expose to training was larger than the pool of officers who would be identified by standard employee monitoring as having engaged in potentially problematic behavior, but smaller than the entire department.

In order to identify our experimental sample, we built on recent literature in criminology which has emphasized the persistence of place in determining highly localized crime rates, commonly known as “hot spots” (Weisburd 2015). A series of studies show that in larger cities about 50% of total crime is located at just 5% of street segments (typically defined as two block faces, curb-cut to curb-cut). In the same way that static features of a small area can contribute to consistently higher crime rates, we hypothesized that particular street segments, or “micro-places,” may pose consistently higher risk for officers being in events which escalate out of control, therefore making it more likely that we may be able to statistically detect some difference in the way trained officers responded to serious citizen encounters. Based on this supposition, we decided that officers working in these “high risk circumstance” micro-places would constitute our sample pool.
Within the city of Seattle, we identified 33,982 micro-places, using street segments as defined by the Washington Department of Transportation. In order to take any spatial correlation into account, we also identified the census block (N=11,512) where the incident occurred. In an extract of all Computer Aided Dispatch (CAD) incidents recorded from January 1st to December 31st 2010, we constructed a history of the micro-place where each incident occurred, based on all other CAD incidents that had occurred in the same street segment and census block over the prior 30 days.\(^{13}\)

In addition, based on our conversations with SPD staff about sources of stress and concern for officers, this history consists of a series of variables describing the frequency and type of incidents in that area, with particular attention to domestic violence, officer safety notices and injuries, people in mental distress, and the ability of dispatchers to provide accurate information about the event to officers. We also included a limited number of officer-specific measures of potential stress measured by the SPD Human Resources Department, specifically a set of binary measures of whether or not an officer was in the 50th percentile of sick days, number of second jobs, or overtime over the past month, consistent with some behavioral models of risk.

We used these histories, described in more detail in Appendix A, to predict whether or not one of three more serious things happened in each event: the CAD incident in question involved force, the officer was injured, or a complaint about the incident (or officer) was filed, using a standard logit model.\(^{14}\) While none of these outcomes intrinsically represents

---

\(^{13}\) We construct each of these nine history measures at the street segment and census block level over the course of the previous month, but also during the same days of the week and time of day of the specific CAD incident. For example, a CAD incident occurring on Saturday at 3 pm would have four street segment history measures: one measure describing the entire month, one describing the previous four Saturdays, one measure describing all previous afternoons, and one describing all previous Saturday afternoons. Four block level history measures were created in the same way.

\(^{14}\) We initially estimated four structural versions of this model: a probit, logit, a skewed logit, and linear probability model. In terms of capturing the relative frequency of complaints, force, and injury across districts in February and March of 2012, a standard logit model fit the data better than the other models. We tested the model using data from February and March of 2012. We did not use data from 2011 because that coincided with the “Occupy Seattle” protests, which almost certainly altered the general spatial patterns and predictors of these
bad policing, all three of them (officers using force, someone using force against an officer, or someone complaining about an officer) are plausibly more likely to be associated with low public perceptions of legitimacy, and also could trigger a standard performance review by the officer’s supervisor. Since, after that review, the supervisors’ may determine that the officer’s behavior was justified by the specific circumstance, we will refer to these outcomes collectively as “potentially problematic” events.

The experiment was active between May and November of 2013. During that time, every two weeks the SPD sent us a list of all CAD incidents during the past 14 days, including information on all officers recorded as responding to each incident. We then compiled the relevant histories for each CAN entry and, using the stored parameters from our 2010 logit model, estimated a predicted risk score for each incident. Each individual officer who was active in CAD was then assigned his or her highest risk score for that two week period. In other words, an individual officer’s predicted risk score was determined by them “checking in” to any CAD event where, in the past, other officers had been involved with the types of incidents that were associated with a higher probability of a potentially problematic event occurring. Officers with higher levels of predicted risk for their precinct were then identified as officers who might be the most likely to be involved in situations where

---

Note that because risk scores were based on past events, the circumstances or outcome of any particular incident was not part of that incident’s predicted risk score. Further, since the risk score assigned to an incident was at best weakly (because of the officer specific HR data) related to an individual officer’s behavior, risk scores could not be manipulated unless the officer stopped working entirely. Even if officers were able to manipulate their risk scores, this manipulation would exclude them from both treatment and control groups, meaning that such manipulation would affect the external rather than internal validity of our estimates.

As was explained in training, there was no way in which having a high predicted risk score should be interpreted by supervisors as a reason for discipline or questioning of that officer’s behavior. All the risk score meant was that, because of where and when this officer worked, they might be more likely, at some point, to disproportionately benefit from additional procedural justice training relative to an officer who worked in areas where incidents had very low predicted risk scores.
additional procedural justice-based supervisory meetings might result in measurable differences in field outcomes.

Every two weeks, we defined officers in the top 12th percentile of predicted risk in each precinct as the sample pool for that period for the reasons that follow. From that sample of officers working in high-risk circumstances, we randomly assigned officers in each precinct to control (no intervention) or treatment (called for supervisory meeting) status for that period’s experimental round. Identifying all officers in the 12th percentile of predicted risk meant that we could a priori expect to maintain a control to treatment ratio of roughly 2:1 while randomly selecting six officers in the two largest precincts to be treated every two weeks, along with two officers from the smallest precinct, and three officers from the remaining precincts. This sample size constraint was imposed due SPD concerns that asking sergeants and lieutenants to conduct any more experimental meetings might be overly burdensome.

As is common in experiments with multiple rounds of randomization (e.g. Heller et. al 2015), officers who appeared in the high risk pool the more than once were randomly re-assigned to control or treatment each time.\(^\text{17}\) As such, our estimates will reflect the average impact of one supervisory meeting – treated and control officers had an equal probability of having participated in a meeting prior to selection, and were equally likely to receive additional training in the future.\(^\text{18}\)

Across all precincts and all periods, the mean predicted risk score for treated officers was very similar to that of control officers; across the treatment groups the mean predicted risk

\(^{17}\)Of all treatment notifications sent, 75\% were sent to officers who had never been selected for an engagement (compared to 79.8\% of control officers), 21.7\% had previously been notified once (13.9\% of control officers), and 3.2 of treated officers (5.4\% of control officers) had been selected twice before. Consistent with randomization, treated officers had been previously notified of being selected for a supervisory meeting 0.008 more times than control officers, which is statistically indistinguishable from zero (p=0.86).

\(^{18}\)If we limit our sample to officers who were eligible for random assignment exactly once, or focus only on officers selected in the first wave of the experiment, our point estimates are essentially identical, although less precise due to the smaller sample size.
was 0.116 (sd =0.186), and across control groups it was 0.120 (sd=0.195). In addition to ensuring that treated and untreated officers are equally likely to benefit from training, and that treatment is not assigned in a way that is dependent on field outcomes, this randomization allows us to disentangle the impact of training on all other events of policy changes affecting police performance by constructing a suitable counterfactual for a treated officer’s behavior over time- a particular concern given the strong seasonality of criminal behavior and criminal opportunities, as well as previously mentioned institutional changes.19

After random assignment occurred, a list of the officers assigned to treatment was submitted to a Seattle Police Department Deputy Chief. These officers and their supervisors were then notified of selection and were instructed, via email, to set up a supervisory meeting, which we referred to as an “engagement,” at the next feasible time, which varied by the work schedules of the officers, sergeants, and lieutenants. For the purposes of evaluation, the “pre” period for each officer consists of all of their activity prior to when these notifications were sent out. Almost all notifications were sent out on Monday morning (some were distributed on Friday). On average, engagements happened 14 days after notification, and 43% occurred within one week after notification emails were sent. There were also a relatively large number of engagements that occurred 11 days after notification, which corresponds to the second Friday after a Monday notification was sent out and the last Friday before the next round of randomization would occur.

Since being notified of selection can be thought of as a separate treatment from being engaged, we consider two “post” periods – officer behavior after notification and behavior after actually having the engagement. There are two reasons why we might expect to see

19 Because of the potential confounding policy changes over time, which are likely as important as individual differences across officers, we introduced an additional eligibility criterion for the treatment and control groups that ensures that the behavior of untreated officers is a true counterfactual for the behavior of treated officers. We identified ten control officers and five treated officers who were on military leave, vacation, or had taken more than five sick days during the week in which they would have been notified of their selection for a meeting, and excluded them from our sample. This ensures that our experimental analysis accurately controls for changes in behavior over time, as these officers had radically different work schedules than all other officers.
different “notification” and “engagement” effects. The first difference is, essentially, the standard intent-to-treat versus treatment-on-the-treated difference, as some varying fraction of officer behavior in the post-notification period is also post-engagement. The second difference is that being told you will have a supervisory meeting and actually having that meeting are, in fact, different treatments; given the nature of most police supervisory meetings, it is reasonable to suspect that officers anticipated the engagements to be punitive in nature despite departmental pre-communications about the study. Comparing outcomes over shorter and longer time periods will provide some evidence on the importance of each mechanism.

4. The Experimental Engagements

Officers selected for treatment were asked to participate in a supervisory training protocol that sergeants could use to reinforce the use of procedural justice by beat officers, developed by the research team in close collaboration with a Deputy Chief. The training protocol that we developed was in many ways a substantial departure from the strictly hierarchical official oversight and training processes used in most police departments, and was inspired by the cognitive interview model (Fisher and Geiselman 2010).

Outside of basic training, police training disproportionately focuses on rare but high-risk (and also high-liability) incidents: use of firearms, use of force, and vehicle operations. Supplemental training, particularly training that requires bringing in outside instructors or sending officers to schools, is dominated by courses dedicated to uncommon scenarios (e.g., hostage situations, edged weapon defense, and active shooter response), and criminal investigation training (e.g., crime scene management and interrogation techniques). There is very little refresher or advanced training on the ”routine” aspects of policing (Walker and Katz 2013, Stoughton 2015).
Departmental supervision tends to reflect an outcome-based approach; supervisors track discrete measures of on-the-job performance (e.g., arrests, traffic accidents, force, and complaints), increasingly through computer-based Early Intervention Systems\textsuperscript{20} or automatic recording devices like in-dash or body-worn cameras. As long as an officer does not use force and is not the subject of a citizen complaint, they are typically given wide latitude in their behavior on the street (Van Mannen 1983). Officers who engage in potentially problematic behavior on the job are subject to formal review and possible discipline, such as forced leave, demotion, termination, or even criminal proceedings. Positive performance, “[making] good, solid, arrests” is rewarded both informally (“respect from non-cynical colleagues and superiors, and the occasional awards banquet”) and formally, through being offered more prestigious work assignments and being targeted for promotion.\textsuperscript{21}

Our supervisory program began when randomly chosen officers were notified of their selection for an engagement. Next, the officer, the officer’s supervisory sergeant, and the lieutenant jointly identified a citizen encounter that the officer had recently participated in based on a selection of the officer’s activity in CAD.\textsuperscript{22} Any incident in which there was an active complaint about the officer, the officer had used force, or that was subject to disciplinary review for some other reason was disqualified. After the incident was selected, the Deputy Chief was notified that the meeting had been scheduled.

Many of the SPD employees were initially confused as to why they were being asked to discuss events where “nothing bad happened” rather than more traumatic things like witnessing a suicide attempt. This response highlights a specific problem implementing

\textsuperscript{20} These systems basically flag officers who engage in more than a predetermined number of these behaviors. In 2003, 29\% of law enforcement agencies surveyed by LEMAS reported using a computer-based monitoring system to identify at-risk officers. Four years later 39\% of surveyed agencies had such a system in place (BJS 2003, 2007). The diffusion of this particular technology is undoubtedly due, at least in part, to incentives put into place by the federal government; the Civil Rights Division of the Department of Justice has recommended the adoption of an EIS program in practically all of its reviews of police department practices. These reports are available online at: http://www.justice.gov/crt/about/spl/findsettle.php#Law%20Enforcement%20Misc

\textsuperscript{21} Communication with Fmr. Deputy Chief, 1/13/2016.

\textsuperscript{22} At the request of the police department, these incidents met at least one of the following criteria: initially classified as “priority one” incident, a report had been filed about the incident, or an arrest was made.
procedural justice training programs; experienced officers respond to “small stuff” all the time, so that they no longer need to actively engage during these events in the same way they did during their first days in the department. The need for additional discussion of these events was similarly unclear to them.23

   At the beginning of the meeting, the supervisory sergeant thanked the officer for attending this supervisory training meeting, and explained that enforcing the law with procedural justice was a central part of the department’s mission and that the goal of that meeting was to discuss how the officer used procedural justice during the incident in question.24 The officer was then invited to discuss the chosen event. Specifically, the sergeant asked the officer to recount how they learned about the incident, what they thought they would see on arrival and what they did observe once they got to the scene.25 Essentially, officers were being prompted to reflect on their thought process during parts of their job where they no longer have to think very carefully—pushing back against their training, and reinforcing the importance of how they made decisions on scene. Did they incorporate new information about the event as it unfolded, or did they essentially switch into autopilot? Our approach accordingly responds to the tension we noted earlier, and pushes officers to think a bit slower in the context of police/citizen encounters.

   As part of this process, sergeants were instructed to ask open ended questions: “What did you observe when you arrived?” versus “Did you notice anything unusual when you arrived?” This put the officer in the position of determining the timing and pace of the meeting - a transfer of control that is atypical for interactions across ranks. Here, the goal was to remind officers that authority does not always perfectly coincide with total control;

23 To some extent, this also reflects a culture in which additional supervision primarily occurs when an officer is facing potential discipline- which can potentially be reflected on the way in which officers approach citizens in the field.
24 At this point, per IRB protocol, the engagement was considered “complete,” and officers were given the opportunity to decide if they wanted to leave. In all cases but one, officers chose to continue the meeting.
25 The script provided to the sergeants is included as Appendix B.
allowing the officer to speak freely did not mean that the sergeant was not in charge of the meeting, in the same way that allowing citizens to speak does not need to diminish an officer’s control over a situation.26

After walking through the (often trivial) events that occurred during the incident, and how the officer responded to the citizen’s actions, the sergeant asked the officer if they were satisfied with the outcome of the incident, and if there was anything that they would have done differently. The sergeants were then instructed to ask the officer for feedback on their performance during the meeting. This query as to how the lower-ranked officer felt about the sergeant’s performance was highly unusual in the policing context, but was intended to model the desired behavior for officers - that they consider how the citizens they encounter feel about their performance on the job.

At the conclusion of the engagement, officers were handed a self-addressed, stamped comment card asking their opinions about the engagement. These anonymous comment cards, addressed to the Police Foundation with postage pre-paid on a return envelope, were intended to be filled out after the meeting, and their responses shielded from view.27

5. Measurement of Outcomes

5.1 Defining Variables

We focused on the impacts of training on what police actually do in the field, specifically on two basic, intuitive, components of successful policing: First, officers should be visible and engaged in the community. They should leave their cars, initiate substantive interactions with citizens, file written reports about any possible concerns that the community has, and arrest individuals who meet a reasonable standard of criminal suspicion. Second, when

---

26 More background on the engagement style can be found in Alpert et al. (2012).
27 Over the course of the experiment we conducted a total of four site visits, observing eight individual sessions, in order to verify that the training protocol was being successfully implemented. Based on observations during site visits, it is clear that the sergeants didn’t always remember to hand these out, and when they did, they failed to write-in the district/precinct as instructed. In addition, some officers filled out only the first side of the two-sided comment card, indicating that supervisors may not have been thoroughly trained in providing this and encouraging its completion.
confronted with potential criminal activity or violations of law, officers should recognize the unique factors that lead to each situation and be flexible in how an incident is resolved, using their best judgment as to what action will maximize social welfare (as opposed to maximizing arrest rates).

In light of these two components, we tested two general hypotheses. First, we examined the impact of training on the frequency and intensity with which officers engage with the public. If training gives officers confidence that they are able to have successful and positive encounters with the community we would expect that, in the days following training, treated officers would be more likely to interact with the community and also spend more time communicating with citizens. Alternately, if being selected for training had the perverse impact of making officers feel “singled out,” we would observe that treated officers were less likely to be involved in CAD incidents, only respond to crimes when directed by dispatch officers, spend less time talking to citizens, and be less likely to file written reports about encounters. This phenomenon of pulling back from the public, particularly in response to increased supervision or scrutiny, is frequently referred to as “de-policing.” Even if every encounter that a police officer has with citizens is positive, de-policing can cause crime rates to increase, people to feel that police have abandoned their community, and worsen public perceptions of police officers (Wolfe and Nix 2015).

We tested this first hypothesis about officer activity using four variables measured in the CAD database. These include the number of CAD incidents the officer was involved in, the fraction of CAD incidents initiated by the officer (“on-views”), the average number of minutes an officer spent on-scene per incident, and the fraction of incidents for which the officer filed a report. Each of these measures is positively correlated with officer activity,

---

28 While we are not aware of any formal evaluation, this type of change (where officers stop initiating encounters with citizens and attempt to limit their time interacting with citizens when they do respond to calls) has been reported as a response to Early Intervention Systems or Early Warning Systems (Amendola 2003).  
29 De-policing in response to increased scrutiny by citizens is increasingly referred to as the “Ferguson effect”
meaning that an increase in all of them suggests an increase in active policing, and a reduction in any of these measures would indicate that additional supervision is associated with de-policing - lower levels of general officer activity and engagement with the community at large.

Our second hypothesis was that officers who receive training would resolve incidents in different ways than otherwise identical officers who do not. In most incidents, officers have substantial discretion as to the outcome of their encounters with citizens. Most do not require any further involvement with the justice system on the part of the citizen. If the officer decides that more serious action is warranted, she or he can issue a ticket or cite the individual. Receiving a citation typically requires someone to pay a fine or appear in court, but does not involve the officer assuming custody of the individual. In more extreme cases, officers can make an arrest, using physical force in the most extraordinary situations. After an encounter, citizens have the option of filing a formal complaint against the officer if they believe they were unfairly treated.

We focused on three main measures of incident outcomes: the fraction of incidents that resulted in an arrest (since this is the disposition that has the most serious criminal justice implications for the involved citizen), the number of times an officer was involved in an incident where physical force was used, and the number of citizen complaints filed against the officer. While a reduction in any of these measures is not a clear policy goal on their own, an excessive number of arrests, uses of force, or complaints relative to the rate at which

---

30 In robustness tests, we will also break down non-arrest outcomes into three categories: citation issued, verbal warning given, or assistance rendered.
31 Whenever an officer uses force to control a citizen, all officers involved in that event are required to complete a use-of-force incident report. Our measure of use of force is therefore based on whether or not an officer filed a use of force report, not whether or not the individual officer actually used force or whether or not the force was justified or deemed consistent with SPD policy, since we do not observe either of these.
32 Like our measure of force, we only observe that a complaint was filed against an officer at a particular date; we do not observe whether or not the complaint was substantiated or not, and we are very limited in our ability to link a complaint to a particular CAD incident given that the people filing the complaint frequently do not report and likely do not know the specific incident number.
officers interact with the community in a more positive manner is cause for concern about the perceived quality of the police force.

Moreover, from the perspective of criminal justice costs, if incidents can be resolved without an arrest, the state is saved the additional costs of criminal justice processing, and the individual and their family does not incur any of the social costs (e.g., eligibility for subsidized housing, financial aid, employment restrictions) associated with potentially having a criminal record.

5.2 Defining Pre- and Post-Test Periods

In order to estimate the counterfactual outcomes for treated offices in each round of randomization, we constructed performance measures from each of the control officers in each engaged officer’s precinct, using notification and treatment dates for the engaged officers. For example, in each round in the Southwest precinct, the six officers who had responded to CAD incidents with the highest risk scores in the previous 14 days were identified as the sample pool in that particular round. Out of this set of six, two were randomly chosen to be engaged, and the remaining four were assigned to the control status in that round.

Suppose that, in round one, both treated officers were notified of their selection on May 1st and, because of scheduling constraints, one treated officer participated in the meeting later that same day but the other was not able to meet with her supervisors until May 8th. This means that the “pre” periods for both of these treated officers are identical, as are the “post-notification” periods, but the “post-engagement” periods for each officer take place at different points in time.

In order to isolate the impact of the engagement from temporal variation in police activity, we essentially construct two observations from the CAD records for each control officer, similar to a “matching with replacement” strategy, using their activity in the weeks
following May 1st as a counterfactual for the first treated officer, and their activity in the weeks following May 8th as a counterfactual for the second. This means that we generate a total of ten “officer-spell” observations from these six officers, where the spells reflect the difference in timing of engagement for the two treated officers – clustering our standard errors by officer to reflect the mechanical correlation in outcomes across different spells for the same control officers.

We will refer to each set of one treated officer and four control officers, all from the same precinct and the same round of randomization, as an experimental block. Our entire sample consists of 221 of these blocks, one for each meeting that took place. Note that within a block, treated and control officers were working in the conditions. Across blocks, however, we would expect geographic and temporal variation in political pressure, police policy, and criminal environment to differentially affect officer behavior.

A-priori, we anticipated that these less than 20 minute engagements would possibly generate immediate changes in behavior that would quickly dissipate. We will therefore focus on outcomes over the one and six weeks before and after treatment. Relative to arrests, uses of force and complaints are rare events, and the low rate at which these occur makes it difficult to interpret the results of traditional statistical analysis. We therefore calculated an even longer run pre- and post- periods for these events: the total number of force incidents and complaints recorded between May 1st and December 31st 2013 for each officer before and after treatment and notification.

In Table 1 we present some descriptive summary statistics for our measures of officer performance, all measured prior to the initial treatment notification dates. There is very little difference across control and treatment groups on any of these dimensions, and in fact the
control and treatment groups are statistically indistinguishable with 95% confidence on all dimensions, implying that randomization was successful.\textsuperscript{33}

We also include a power calculation in Table 1, where we estimate the probability of rejecting the null hypothesis of no treatment effect with 90% certainty if the engagements actually cause a 10% change in officer behavior. Our tests of officer activity have high levels of power, almost all over 80%. Our tests for incident response have lower levels of power, meaning that a failure to reject a null hypothesis should, for the most part, be interpreted as a failure to draw any conclusions, rather than evidence of a null effect.

On average, officers responded to 40 CAD events in the one week pre-period before they were notified of their engagement. Of those incidents, 32% were initiated by the officer, with an average time of 40 minutes from initial dispatch to return to service. On average, 30% of incidents were serious enough for the officer to file a report, and 6.5% resulted in an arrest being made. Most incidents are resolved by the officer simply rendering assistance. In a given week, 0.003 complaints are filed against each officer, and each officer files 0.2 reports about being involved in a use of force incident. These pre-treatment means, along with the low level of power, lead us to heavily discount one week outcomes for these last two measures; in our analysis of force and complaints, we will focus on six week outcomes, along with a comparison of the longest possible pre and post periods for each group.

When we examined the six week pre-periods before officers were notified of their selection, we found that officers are involved in more CAD events, have more complaints filed against them, and are involved in more incidents where force is used. This is what we should see, since we are looking at a longer time period. Turning to the measures of typical

\textsuperscript{33} This was statistically verified by re-estimating 20 modified versions of our central outcome equation, where the dependent variable was PreOutcome\textsubscript{ijw} for each of our treatment windows and all outcomes. None of our pre-notification outcomes are statistically distinguishable with 95% confidence, and only one (CAD events resolved by issuing a non-criminal citation over one week period) was statistically precise at the 90% level of confidence (p=0.083).
incident outcomes for each officer, we also observe little difference across control and treatment groups.

6. Evaluation Model

In order to identify the impact of the meetings on officer behavior, we must adjust the comparison of control and treatment outcomes to reflect the block level randomization, and for the fact that each individual officer could appear in multiple blocks. As a first pass, we can incorporate each of these features by estimating as equation 1:

Eq. 1: \( PostOutcome_{ijb} = \alpha_{jb} + \beta \text{Engaged}_{ijb} + \varepsilon_{ijb} \)

where \( PostOutcome_{ijb} \) measures the behavioral outcomes (measured over one or six weeks) for officer \( i \), in jurisdiction \( j \) in block \( b \). The block specific intercept \( \alpha_{jb} \) subtracts out the mean value of outcomes for the treated and control officers in the same block, meaning that any department-wide or precinct-specific events that might influence officer activity are differenced out of our results. \( \text{Engaged}_{ijb} \) is a dummy variable equal to one if officer \( i \) was treated in that particular block, and \( \varepsilon_{ijb} \) represents the remaining unexplained component of the behavioral outcome. By construction \( \varepsilon_{ijb} \) is uncorrelated with \( \text{Engaged}_{ijb} \), but will be correlated for observations from the same officer, so we cluster our standard errors at the officer level. Our estimate of \( \hat{\beta} \) is defined as the average difference in outcomes for the treated officer from control officers in her block, who were working in her precinct, in similarly risky areas, at the same points in time, but were not randomly selected to have an engagement at that time.

Since we did not identify any systematic differences in our measures of officer behavior across treated and control groups prior to randomization, \( \hat{\beta} \) from this model can be interpreted as an unbiased estimate of the impact of having a supervisory meeting on the eligible officers. Of course, it is possible that treated officers may have discussed the engagements with their colleagues, who may or may not be in the control group in a given
time period. This contamination effect would tend to bias our results towards zero, meaning our estimates should be interpreted as plausible lower bounds of the impact of training.

Two extensions to the model can increase its explanatory power, reducing the amount of noise in our measures of behavior and allowing for more precision in our estimates. First, we can include a control for an officer’s behavior in the one or six weeks prior to notification, as in the lagged dependent variable model in equation 2:

Eq. 2: \( \text{PostOutcome}_{ijb} = \alpha_{jb} + \theta \text{PreOutcome}_{ijb} + \beta \text{Engaged}_{ijb} + u_{ijb} \)

Equation 2 assumes that there is one unknown parameter \( \theta \) that characterizes the linear relationship between officer \( i \)'s behavior prior to notification and afterwards, and this relationship is common across control and treatment observations. Alternately, we can fix the relationship between pre- and post-treatment behavior for each officer-spell, but allow this fixed relationship to vary by precinct and over time by using a first-differences model, as in equation 3:

Eq. 3: \( \text{PostOutcome}_{ijb} - \text{PreOutcome}_{ijb} = \alpha_{jb} + \beta \text{Engaged}_{ijb} + v_{ijb} \)

These models are not identical; they are based on different assumptions about the nature of the “noise” in an officer’s behavior—does it follow an AR(1) process, or is it time invariant? More to the point, whether or not equation 2 or equation 3 is more appropriate depends on the source of unexplained variation in post-treatment outcome, and whether or not it is time invariant (e.g. some officers always use a lot of force) or something more dynamic (e.g. officers who lots of force at one point in time may be reassigned to beats where they will use less force in the next period). Based on our observations of officers and discussions with the Deputy Chief, we tend to think that a substantial amount of behavioral variation across officers is time invariant (equation 3) rather than AR(1) (equation 2), but we will present results for all three models as a way of bounding the true treatment effect size.

7. **Results: Did Meetings Affect On-the-Job Performance?**
7.1 Officer Activity and Community Engagement

If officers who were selected for meetings felt singled out, it is possible that they could try to avoid future meetings by de-policing. Alternately, reviewing a recent citizen encounter with their supervisors could provide officers with additional confidence in their ability to interact with citizens, and lead them to increase their engagement with the community.

< Table 2 about here >

Table 2 displays our estimates of the impact of supervisory meetings on officer activity measures one and six weeks before and after treatment, for all three models. We found no evidence that treated officers reduced their activity in response to being notified of or undergoing an engagement. In fact, these estimates suggest that, relative to otherwise similar officers, treated officers participated in roughly 3 to 4 more incidents in the week after being notified or treated (a 10% increase in activity). Perhaps in anticipation of the meeting, the officers were more likely to initiate the interaction and spend more time on scene. When the analysis period is extended to six weeks before and after the meetings, there is less evidence of a change in community engagement. Not only are the point estimates relatively imprecise, but both coefficient and standard error estimates are very small relative to the pre-treatment mean over this longer term. There is also little substantively or statistically important difference in the magnitude of estimates from the three different models.

Of course, there are other ways in which police officers could engage in de-policing that could be elicited through other methods, such as detailed surveys. However, in terms of actual data that the SPD collects and official metrics used to publicly characterize community engagement, we find no evidence of an adverse effect of these meetings, and if anything observe an increase in the presence of officers in the community. This is a markedly different finding from anecdotes of “de-policing” in response to increased scrutiny of police officers.
(e.g. Comey 2015). In addition, the possible increase in officer engagement with the community could be associated with increased deterrence (e.g. Owens 2013).

< Table 3 about here >

7.2 Incident Outcomes

Almost all communities want police officers to be a visible presence in their neighborhood, but not if those officers appear to arrest local residents indiscriminately or use force against them without clear cause. Our estimates of how supervisory meetings affected the ways in which officers interacted with the public are presented in Table 3. While we failed to find evidence of de-policing on the part of officers, we do find evidence that officers who have non-disciplinary supervisory meetings are more likely to resolve those encounters in non-disciplinary ways.

We do not observe a systematic change in the number of arrests per CAD incident made by officers one week after they are notified that they will be selected. However, these same officers are roughly 25% less likely to decide to make an arrest in the week after the meeting actually occurs. This effect is consistent across all three models, and in absolute terms, this corresponds with a reduction in the probability of arrest from 6% to 4.5%. Of course, we cannot say with certainty that reminding officers to resist automation on the job leads them to conclude that fewer arrests are necessary, rather than officers choosing to make fewer arrests (but, recalling the results in table 1, simultaneously choosing to interact with the public at least as much) in order to avoid further training. However, if officers believed that the meetings would be punitive, and were simply trying to avoid being part of our sample in the future, we might expect to see this response after being notified, rather than after being trained.

When we compare how treated and control officers resolve CAD incidents over the six weeks before and after the engagements, we continue to observe a reduction in the rate at
which officers resolve incidents in punitive ways; treated officers are roughly 12 percent
(p=0.06) less likely to make arrests after a supervisory meeting. Again, assumptions about
unobserved heterogeneity in outcomes do not appear to influence the relationship between
officer training and their arresting behavior.

Perhaps even more so than unnecessary arrests, allegations of excessive force by officers
can have dire consequences for perceived police fairness and effectiveness. Since less than
0.05 percent of CAD incidents involve force or a citizen complaint, we focus on longer time
periods for these outcomes. During the six-week period after notification and engagement, we
estimate that engaged officers are between 16 and 50 percent less likely to be involved in
force incidents than control officers, and there is reasonable likelihood (at least a 13 percent
chance) that a reduction of this size would be observed by chance. While these estimates are
not statistically significant by conventional standards, given the importance of force in officer
and community relationships, and the relatively low cost of this intervention, we consider this
result highly promising, and worthy of further experimental analysis. The magnitude of this
change is roughly equivalent to the current best estimate of the impact of body-worn cameras
on use of force (Ariel et al. 2015).

In the next columns of Table 3 we expand our time frame to include the entire sample
period, between May and December of 2013. Essentially, we are now comparing the
temporal distribution of force events across control and treatment groups, and determining
whether or not treated officers are less likely to be involved in force incidents at any time
after their supervisory meetings than they were prior to the meeting. Roughly one out of
every two officers who work in the highest risk circumstances are involved in events with
force prior to engagement, and we estimate that there is a roughly 15 to 40 percent reduction
in the likelihood that treated officers are involved in these potentially problematic
situations. The precision of our estimates suggests that it is less likely to observe this pattern of behavior simply as a matter of chance, with p-values ranging from 19 to 2 percent.

In contrast to our arrest and force results, we find little evidence that engaged officers have complaints filed against them at different rates than officers in the control groups. All of our point estimates are positive, but also noisy; looking across models, we cannot reject the null hypothesis that the meetings increased complaints by more than 90 percent or reduced them by over 70 percent with 95 percent certainty. Therefore, we must conclude that the available evidence does not allow us to say anything about whether or not these additional meetings have any effect on the frequency with which citizens file complaints against SPD officers.

7.3 Are Officers Systematically Resolving Incidents in Other Ways?

If officers are systematically resolving incidents that previously would have resulted in an arrest in one particular way, then we might observe an increase in the fraction of CAD events resolved by issuing a non-criminal citation, a warning, or rendering assistance. Of course,

---

34 What our results tell us is that officers who have had supervisory meetings are less likely to be involved in an incident where any of the involved officers feel that physical force must be used to regain control of the situation. Only one officer using force may cause multiple officers to file force reports, but if any one officer is able to maintain control of a situation without resorting to physical force, it is plausible that no force will be used at all. The impact of this measurement error means that, if anything, our results are underestimates of the true effect.

35 Our previous measures of officer activity and incident outcomes were all continuous variables which primarily ranged from zero to one. Use of force is a much more restricted variable that takes on one of six values. If the observed average reduction in force was driven by one officer who was involved in six incidents prior to engagement and none afterwards, this might limit the interpretability of our results. In Appendix C, we plot the pre-post difference in involvement in force incidents for each observation in our sample. This figure suggests that, in fact, our identification of a reduction in force incidents appears to be driven by multiple treated officers engaging in roughly one less use of force incident after engagement, rather than one treated officer making a large change. We further confirmed that outliers are not driving our results by replicating Table 3 using a “trimmed” sample of force incidents. In this sample, any officer who was involved in three or more incidents where force was used was recorded as being involved in exactly three force incidents. These results, available on request, are essentially identical to the full sample.

36 It is possible that complaints that happened prior to notification are complained about in the “post” period. This sort of measurement error should affect both our control and treatment groups equally, and therefore not affect our experimental estimates. It is also possible that citizens will file complaints about incidents that occurred in the “post” period after our sample window. If non-disciplinary supervisory meetings affected officer behavior in a way that changed the amount of time it took for citizens to file complaints, this type of measurement error would affect our estimates. As our sample window gets larger, this type of measurement error is less likely to be a problem, as we will be allowing for a longer period of time between incident and complaint.
given that arrests are so infrequent (roughly seven percent of all incidents), a large reduction in arrests may not produce a substantively meaningful increase in any other non-arrest outcome.

< Table 4 about here >

Our examination of alternate outcomes is presented in Table 4. We find some evidence that, in the short run, officers are four percentage points more likely to be involved in incidents are cleared as “assistance rendered,” which is the primary way in which officers resolve interactions, and involves no punitive consequences for the citizen. It is also notable that, while only marginally precise in one model, the magnitude of the increase in citations is very similar to the observed reduction in arrests. From a citizen’s perspective, receiving a citation means that one must pay a fine or appear in court to contest the citation, but you are not immediately held in jail and it is highly unlikely that you will accrue a criminal record.

7.4 Shorter Term Outcomes

A priori, we anticipated that officers would respond to these brief interventions by immediately changing their behavior, but that the effects of a short supervisory meeting would dissipate quickly. Further, because most engagements occurred 11 days after notification, the comparing shorter term outcomes after notification and engagement can help us identify changes associated with the anticipation of a meeting versus changes occurring after the engagement. Only 3.2% of engagements had taken place in the 24 hours after notification, 6.8% took place within 48 hours, 42.1% within one week, 91% within one month of notification, and 95.9% of engagements took place within six weeks. In figures 1 - 10, we graphically present standardized estimates for the impact of a supervisory meeting on officer behavior in 24 hour, 48 hour, one week, two week, one month, and six week intervals after being notified or engaged. Any behavioral change in the first 48 hours after being notified can be roughly interpreted as a “notification effect,” and by four weeks, essentially
all control officers have actually had the engagement. All treatment effects are scaled by the control pre-treatment means so they can be interpreted as percent changes, and we also present upper and lower bounds of 90% confidence intervals.

The first thing immediately apparent from the graphs is that telling officers they have been selected for a supervisory meeting immediately increases their activity relative to the control group, and this effect lasts for at least 48 hours. Officers also are generally more active immediately after the engagements take place (figure 1); they are more likely to initiate citizen encounters (figure 2), they spend more time on the scene, (figure 3) and they are more likely to file a report after being told they will have a meeting, or when they have a meeting, relative to the 48 hours prior to being notified (figure 4). It is also noteworthy that, in models where unobserved individual heterogeneity is modeled as an AR(1) process, officers are more likely to make arrests (as well as choose any of the other three most common dispositions) in the 24 hours after being notified that they will meet with their supervisors. However, after actually having the meeting, this increased propensity to arrest citizens quickly disappears, and on average, treated officers are consistently less likely to resolve encounters with an arrest in the following weeks.

With the exception of one week outcomes, which have essentially a zero base rate, we observe a relatively consistent reduction in the officer’s involvement in incidents where force is used (figure 6). The precision of the estimates increases as the length of the pre and post periods increases. Citizen complaints also appear to be consistently slightly higher for the treated group, and always imprecise when the primary source of unobserved variation across officer behavior is assumed to be time invariant (model 3).

Consistent with table 4, we find that in the 48 hours following an engagement, officers are more likely to issue citations (figure 8) rather than make arrests, an effect which is consistent across all models. Recall that citations are second to arrest in the sense of formal
criminal justice involvement – citations typically require someone to pay a fine, or potentially appear in court at a later date, but the officer doesn’t actually take you into custody. We observe only imprecise changes in the frequency with which officers report that they issued a verbal warning (figure 9), but a short run increase in the rate at which officers simply describe the result of an encounter as “assistance rendered” (figure 10), particularly during the 24 hours after they are notified that they will receive additional supervision in the future.

One possible interpretation of the pattern of results is that treated officers choose lesser sanctions, rather than completely ignoring criminal behavior. Whether or not this “downgrading” of officer resolution from arrest to citation is socially optimal depends, of course, on whether or not the officers were previously “warriors” who were making excessive arrests. We will return to this question in section 7.6.

7.5 Effect Heterogeneity by Officer Risk

There are some circumstances where officers will always make arrests or always use force and some circumstances where neither of those two things would ever happen. Additional training in how officers use information on scene will have the largest impact of incident outcomes in cases where interactions may, or may not, become problematic, depending on the officer’s skill in handling the situation. Because all of the officers in our sample worked on beats with elevated levels of predicted risk, officers with lower predicted risk scores may be more likely to be on that behavioral margin.

In Table 5, we include the first order impact of the predicted risk score that qualified the particular officer for selection into the sample (as treatment or control), and an interaction of risk score and selection for treatment. To ease interpretation, risk scores are standardized to have a mean of zero and a standard deviation of one.

< Table 5 about here >
Logically, the marginal officers eligible for meetings should work in less relatively risky places, and therefore be more likely to be in situations where an officer’s use of cognitive capital matters for the outcome of an interaction. Our data are consistent with this; we find that officers who work in areas with higher predicted risk scores have lower reductions in arrest rates. On average, engaged officers are 0.7 percentage points (roughly 10%) less likely to resolve an incident via arrest. However, treated officers who work in areas that are one standard deviation higher in predicted risk than average do not behave differently than control officers. Officers who work in areas that are one standard deviation less risky are 20 percent less likely to resolve incidents via arrest than their otherwise similar colleagues. Since we might expect cognitive capital engagement to affect marginal situations, this is also more consistent with our central hypothesized mechanism than deterrence, where we might expect officers to be less likely to make arrests in all situations to simply avoid future meetings.\(^{37}\)

### 7.6 How Costly Were the Meetings? Officer Surveys Results and Overall Crime

At the end of each supervisory meeting, officers were given pre-addressed, postage paid survey cards, which could be mailed directly to the Police Foundation.\(^{38}\) The surveys consisted of a combination of closed form and open ended questions intended to elicit the officer’s perceptions of the meeting. In addition to providing a test of implementation fidelity, the response to the officers provides us with a sense of how disruptive and costly

---

\(^{37}\) A priori, we anticipated two sources of potential heterogeneity in the impact of these supervisory meetings: the number of meetings an officer was engaged in and the level of predicted risk where they worked. Ex post, 33 of the 42 officers who were randomly selected to participate in two meetings worked in either the North or West precincts (where 131 of all 221 meetings took place), and only seven officers were selected to participate in three meetings. This concentration of multiple meetings in a few precincts makes it difficult to interpret any heterogeneity in the effect of these meetings along this dimension; F tests of the joint significance of the interactions between the 76 supervisory sergeants and the treatment effect suggest that there was variation in the impact of meetings based on who lead them (F- statistics ranging from 9 to 98, depending on outcome, all corresponding to p-values of less than 0.001). We also estimated models where we allowed for different treatment effects by precinct and by round of randomization, and found no obvious patterns.

\(^{38}\) Out of a possible 221 cards distributed, a total of 66 surveys were received by the Police Foundation, for a maximum response rate of 29.9 percent. Since the surveys were anonymous, we cannot tell how many unique officer opinions are represented by these responses, or if the responding officer participated in more than one meeting. It was not clear if the low response rate reflected an unwillingness to participate, as the responses were voluntary, or if supervisors had not fully followed through at the end of the training. In some cases where cards were returned, only one side was completed, indicating incomplete instructions by supervisors.
these interventions were; if officers found the meetings to be intrusive and onerous, a
continuation of this type of intervention would only be justified if there were large socially
desirable behavioral effects. In addition, we would be more concerned about officers
changing their behavior to avoid meetings if these meetings were viewed as threatening or
unpleasant.

< Table 7 about here >

As shown in Table 7, officers responded to the meetings in a generally positive way,
consistent with the goals of procedural justice. Officers appeared to most consistently feel as
if what they had to say mattered to the supervisor, and they were not personally or
professionally judged for their opinions. Notably, and consistent with the moderate level of
confusion about the benefit of discussing “small stuff” with their supervisors, officers
reported neutral feelings about whether or not they felt “more respected” than previous
meetings, and whether or not they anticipated these meetings would change their
performance in the field. The most common comment was that officers simply wanted to talk
about more “exciting” events (n=15). Taken as a whole, these results suggest that these
meetings were viewed as relatively innocuous, and at worse a trivial distraction from typical
duties. The open ended responses suggest that providing officers an opportunity to interact
with sergeants in a less authoritarian manner, and in a format without an obvious disciplinary
goal, was viewed as a departure from standard practice.

Finally, the social cost of the training program could be quite high if crime rose in
response to the reduction in arrests. Figure 11 presents monthly index crime rates, excluding
murder, in Seattle, Portland Oregon, and Spokane Washington from the 2010 to 2014
Uniform Crime Reports. Crime rates and crime trends appear to be generally similar across
these three cities, with the exception of car theft, which seems to start rising in the beginning
of 2013 in Seattle, and continues to rise after the experiment was over.
None of the figures reveal anything exceptional about crime in Seattle during or immediately following the implementation of the engagements. Of course, since only 180 officers (just under 14% of the sworn force) were treated, and control and treated officers worked in the same places by construction, these results should be interpreted with caution, as it is possible that a larger scale program could yield very different conclusions. However, the available data provide no evidence that the avoided arrests and lower rates of force due to this training program had any substantial impact on overall crime in Seattle. Further, quasi-experimental evidence suggests that it is police officer presence, rather than arrests per se, that reduce crime (Owens 2013).

8. Conclusion

How do police departments convince citizens that they are simultaneously guardians who protect and serve and crime-fighting warriors who make arrests that can lead to costly sanctions? In cooperation with the city of Seattle and the Seattle Police Department, we developed a low-intensity supervisory program for police officers aimed at altering the way they interact with the public. The goal of this training was to slow down officer decision making in low-stress, common, citizen encounters. We identified at the outset a tension between the routinized everyday response to police/citizen encounters that comes with police experience and the slower thinking needed to carry out procedural justice enhanced policing. This is a tension that prior studies have not explicitly recognized, and our intervention was developed to respond directly to this issue. In addition, our intervention highlighted a second tension between the goals of police reformers and the realities of police work; the hierarchical nature of officer supervision does not provide a model for how officers could interact with citizens in a way that is procedurally just.

Officers randomly selected to participate in these supervisory meetings were as active in the community as a control group, but were 12 to 25% less likely to resolve incidents with an
arrest and were up to 60% less likely to be involved in use of force incidents. We found that
the largest reduction in arrests occurred among officers who were more likely to be working
in areas where there was only a moderate probability of being in a "risky circumstance”

This is an important finding in the context of continuing concerns that the police are
unnecessarily using their enforcement powers in a way that alienates citizens from the police
(Rosenbaum et al. 2005). Indeed, our estimated behavioral changes are roughly in line with
the one experimental evaluation of the impact of body-worn cameras on police officer
behavior, which was conducted in a small department of roughly 50 officers (Ariel et al.
2015). Overall, we conclude that non-disciplinary, procedural justice based supervisory
meetings are a promising strategy for improving police-community relations specifically and
policing in general.

As with any experiment, there are important limitations to this study and caveats in
interpretation. For many of our outcomes, particularly complaints, use of force, and non-
arrest outcomes, we have a low level of statistical power, which is not surprising due to their
rarity. In addition, the estimated rate of type 1 error in some of our estimates was between 5
and 10 percent. Of course, from a policy standpoint, the acceptable level of type 1 error is a
function of the cost of program implementation and the potential for adverse effects. Given
that we found no evidence of officers responding to meetings by de-policing and that this was
not a costly intervention in terms of officer perceptions of supervisor time, p-values of less
than 10 percent strike us as sufficient to conclude that these types of proactive, non-punitive
supervisory meetings hold promise. We also found no evidence of a differential change in
crime in Seattle during our experiment, although this finding may not hold in a situation
where all police officers, rather than a subset, were treated, or in cities with fundamentally
different criminal environments.
Only limited attention has been paid to the causal relationship between changes in policies governing police behavior and public perceptions of police integrity.\textsuperscript{39} There are also very few, if any, experimental evaluations of police training programs on real world outcomes. The goals of our training are based on the idea of procedural justice, and the aim of the training was to change specific officer behaviors that cause public dissatisfaction with the police: arrest rates that are too high and use of force that is believed to be unjustified or result in disparate outcomes for members of minority groups. If supervisory tactics that reduce the rate at which officers engage in exactly these behaviors are not associated with improvements in perceived police legitimacy, this may be a signal that change in criminal justice policy or guidelines for officer behavior is limited in its ability to improve police quality.

\textsuperscript{39} Heaton (2010) is a notable exception. See Johnson et al. (2014) for a detailed discussion of this issue.
References


Rosenbaum, Dennis P., and Daniel Lawrence. 2011. Teaching respectful police-citizen encounters and good decision making: Results of a randomized control trial with police
http://www.nationalpoliceresearch.org/the-chicago-quality-interactio/


Weisburd, David. 2015. The law of crime concentration and the criminology of place. 
Willis, James. 2013. Improving police: What’s craft got to do with it? Monograph of lecture 
presented at the Police Foundation’s Ideas in American Policing series, Washington DC: 
Police Foundation Press.
University Press.
Yamaguchi, Motonori, and Robert W. Proctor. 2011. Automaticity without extensive 
training: The role of memory retrieval in implementation of task-defined rules. 
Figures

Figure 1: CAD Incidents per Officer: Effect Sizes and 90% Confidence Intervals

Notes: Black Triangle: Model 1, Black X: Model 2, Grey Line: Model 3

Figure 2: Percent On-Views: Effect Sizes and 90% Confidence Intervals

Notes: Black Triangle: Model 1, Black X: Model 2, Grey Line: Model 3
Figure 3: Minutes On Scene: Effect Sizes and 90% Confidence Intervals

Notes: Black Triangle: Model 1, Black X: Model 2, Grey Line: Model 3

Figure 4: Written Reports per Incident: Effect Sizes and 90% Confidence Intervals

Notes: Black Triangle: Model 1, Black X: Model 2, Grey Line: Model 3
Figure 5: Arrests per Incident: Effect Sizes and 90% Confidence Intervals

Notes: Black Triangle: Model 1, Black X: Model 2, Grey Line: Model 3

Figure 6: Use of Force: Effect Sizes and 90% Confidence Intervals

Notes: Black Triangle: Model 1, Black X: Model 2, Grey Line: Model 3
Figure 7: Citizen Complaints Filed: Effect Sizes and 90% Confidence Intervals

Notes: Black Triangle: Model 1, Black X: Model 2, Grey Line: Model 3

Figure 8: Citations Issued per Incident: Effect Sizes and 90% Confidence Intervals

Notes: Black Triangle: Model 1, Black X: Model 2, Grey Line: Model 3
Figure 9: Warnings Issued per Incident: Effect Sizes and 90% Confidence Intervals

Notes: Black Triangle: Model 1, Black X: Model 2, Grey Line: Model 3

Figure 10: Assistance per Incident: Effect Sizes and 90% Confidence Intervals

Notes: Black Triangle: Model 1, Black X: Model 2, Grey Line: Model 3
Figure 11: Monthly Crime Rates in Seattle, Portland OR, and Spokane WA, 2010-2014

Notes: Black line: Seattle PD, Medium Gray line: Portland PD, Light Gray line: Spokane PD. UCR Index Crimes per 1,000 residents. Timing of experiment indicated by vertical lines.
<table>
<thead>
<tr>
<th>Activity Measures</th>
<th>One Week Prior</th>
<th>Six Weeks Prior</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Treated</td>
<td>Control</td>
</tr>
<tr>
<td></td>
<td>N=221</td>
<td>N=1213</td>
</tr>
<tr>
<td>Incidents</td>
<td>43.33</td>
<td>40.32</td>
</tr>
<tr>
<td></td>
<td>(18.48)</td>
<td>(22.90)</td>
</tr>
<tr>
<td>Report Taken</td>
<td>0.299</td>
<td>0.283</td>
</tr>
<tr>
<td></td>
<td>(0.128)</td>
<td>(0.158)</td>
</tr>
<tr>
<td>On Views</td>
<td>0.318</td>
<td>0.325</td>
</tr>
<tr>
<td></td>
<td>(0.198)</td>
<td>(0.223)</td>
</tr>
<tr>
<td>Minutes On Scene</td>
<td>39.61</td>
<td>39.74</td>
</tr>
<tr>
<td></td>
<td>(15.85)</td>
<td>(23.57)</td>
</tr>
<tr>
<td>Incident Outcomes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arrests</td>
<td>0.070</td>
<td>0.060</td>
</tr>
<tr>
<td></td>
<td>(0.071)</td>
<td>(0.077)</td>
</tr>
<tr>
<td>Force Reports (total)</td>
<td>0.018</td>
<td>0.018</td>
</tr>
<tr>
<td></td>
<td>(0.134)</td>
<td>(0.134)</td>
</tr>
<tr>
<td>Number of Complaints (total)</td>
<td>0.00</td>
<td>0.004</td>
</tr>
<tr>
<td></td>
<td>(0.00)</td>
<td>(0.064)</td>
</tr>
<tr>
<td>Citations</td>
<td>0.239</td>
<td>0.224</td>
</tr>
<tr>
<td></td>
<td>(0.117)</td>
<td>(0.131)</td>
</tr>
<tr>
<td>Warnings</td>
<td>0.051</td>
<td>0.052</td>
</tr>
<tr>
<td></td>
<td>(0.061)</td>
<td>(0.080)</td>
</tr>
<tr>
<td>Assistance Rendered</td>
<td>0.489</td>
<td>0.474</td>
</tr>
<tr>
<td></td>
<td>(0.139)</td>
<td>(0.181)</td>
</tr>
</tbody>
</table>

Standard Deviations in Parentheses.
* a standard error of 0.001 is assumed for the treatment group
Table 2: Supervisory Meetings and Officer Activity (n=1,434)

<table>
<thead>
<tr>
<th>Model</th>
<th>One Week</th>
<th>After Notification</th>
<th>After Meeting</th>
<th>After Notification</th>
<th>After Meeting</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(a)</td>
<td>(b)</td>
<td>(c)</td>
<td>(a)</td>
<td>(b)</td>
</tr>
<tr>
<td>Total CAD Events</td>
<td>5.12**</td>
<td>3.86*</td>
<td>2.85</td>
<td>4.42*</td>
<td>3.48*</td>
</tr>
<tr>
<td></td>
<td>[1.85]</td>
<td>[1.61]</td>
<td>[1.82]</td>
<td>[1.71]</td>
<td>[1.55]</td>
</tr>
<tr>
<td>P(Effect = 0)</td>
<td>0.006</td>
<td>0.0173</td>
<td>0.119</td>
<td>0.0101</td>
<td>0.0258</td>
</tr>
<tr>
<td>Control Mean</td>
<td>40.32</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R²</td>
<td>0.17</td>
<td>0.39</td>
<td>0.16</td>
<td>0.25</td>
<td>0.38</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>“On-Views”</td>
<td>0.014</td>
<td>0.024+</td>
<td>0.027+</td>
<td>-0.0092</td>
<td>-0.0011</td>
</tr>
<tr>
<td></td>
<td>[0.017]</td>
<td>[0.013]</td>
<td>[0.014]</td>
<td>[0.015]</td>
<td>[0.011]</td>
</tr>
<tr>
<td>P(Effect = 0)</td>
<td>0.422</td>
<td>0.0751</td>
<td>0.0545</td>
<td>0.540</td>
<td>0.923</td>
</tr>
<tr>
<td>Control Mean</td>
<td>0.325</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R²</td>
<td>0.25</td>
<td>0.6</td>
<td>0.22</td>
<td>0.29</td>
<td>0.53</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time on Scene</td>
<td>3.02</td>
<td>3.04</td>
<td>3.08</td>
<td>1.47</td>
<td>1.49</td>
</tr>
<tr>
<td></td>
<td>[2.89]</td>
<td>[2.73]</td>
<td>[2.75]</td>
<td>[1.95]</td>
<td>[1.85]</td>
</tr>
<tr>
<td>P(Effect = 0)</td>
<td>0.297</td>
<td>0.265</td>
<td>0.264</td>
<td>0.451</td>
<td>0.421</td>
</tr>
<tr>
<td>Control Mean</td>
<td>39.74</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R²</td>
<td>0.19</td>
<td>0.25</td>
<td>0.17</td>
<td>0.22</td>
<td>0.29</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Written Report Filed</td>
<td>0.014</td>
<td>0.0096</td>
<td>-0.0019</td>
<td>0.012</td>
<td>0.0055</td>
</tr>
<tr>
<td></td>
<td>[0.013]</td>
<td>[0.012]</td>
<td>[0.014]</td>
<td>[0.012]</td>
<td>[0.012]</td>
</tr>
<tr>
<td>P(Effect = 0)</td>
<td>0.250</td>
<td>0.432</td>
<td>0.895</td>
<td>0.318</td>
<td>0.635</td>
</tr>
<tr>
<td>Control Mean</td>
<td>0.283</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R²</td>
<td>0.23</td>
<td>0.29</td>
<td>0.16</td>
<td>0.26</td>
<td>0.36</td>
</tr>
</tbody>
</table>

Notes: The dependent variable in models (a) and (b) is the level outcome in the post period. Model (b) includes controls for the pre-Notification period outcome in levels. In model (c) the dependent variable is the linear difference in outcomes relative to pre-Notification period for each officer. All regressions include experimental block fixed effects. Robust standard errors in outcomes allow for arbitrary correlation within officer (320 clusters). The reported control mean is calculated in the pre-notification period.

+ p < .10  * p < .05  ** p < .01  *** p < .001
Table 3: Supervisory Meetings and Incident Outcomes (n=1,434)

<table>
<thead>
<tr>
<th>Model</th>
<th>One Week After Notification</th>
<th>After Meeting</th>
<th>Six Weeks After Notification</th>
<th>After Meeting</th>
<th>Six Weeks Full Time Period</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(a)</td>
<td>(b)</td>
<td>(c)</td>
<td>(a)</td>
<td>(b)</td>
</tr>
<tr>
<td>Fraction of Events Resulting in Arrest</td>
<td>0.0035</td>
<td>0.0011</td>
<td>-0.0011</td>
<td>-0.011+</td>
<td>-0.014*</td>
</tr>
<tr>
<td></td>
<td>[0.006]</td>
<td>[0.005]</td>
<td>[0.006]</td>
<td>[0.006]</td>
<td>[0.006]</td>
</tr>
<tr>
<td>P(Effect = 0) Pre-Notification Control Mean</td>
<td>0.523</td>
<td>0.837</td>
<td>0.862</td>
<td>0.0799</td>
<td>0.0163</td>
</tr>
<tr>
<td>R²</td>
<td>0.18</td>
<td>0.41</td>
<td>0.25</td>
<td>0.18</td>
<td>0.38</td>
</tr>
</tbody>
</table>

| Total Use of Force Incidents | -0.0354 | -0.0364 | -0.0738 | -0.0241 | -0.0236 | -0.0625 | -0.0792 | -0.088+ | -0.207* | -0.0596 | -0.0644 | -0.187* |
|                            | [0.035]  | [0.035]  | [0.049] | [0.027] | [0.027] | [0.049] | [0.0536] | [0.0524] | [0.0890] | [0.0460] | [0.0455] | [0.0904] |
| P(Effect = 0) Pre-Notification Control Mean | 0.309 | 0.296 | 0.13 | 0.377 | 0.385 | 0.202 | 0.141 | 0.094 | 0.021 | 0.196 | 0.158 | 0.0391 |
| R²       | 0.216                        | 0.217        | 0.166                       | 0.251        | 0.251        | 0.184        | 0.312   | 0.317   | 0.366   | 0.321   | 0.323   | 0.363   |

| Total Citizen Complaints | 0.0099 | 0.0098 | 0.0122 | 0.028+ | 0.028+ | 0.0307 | 0.0194 | 0.0195 | 0.0128 | 0.0266 | 0.0267 | 0.02 |
|                         | [0.016] | [0.016] | [0.032] | [0.016] | [0.016] | [0.032] | [0.0294] | [0.0293] | [0.0483] | [0.0287] | [0.0287] | [0.0478] |
| P(Effect = 0) Pre-Notification Control Mean | 0.532 | 0.533 | 0.7 | 0.0854 | 0.0853 | 0.337 | 0.51 | 0.506 | 0.791 | 0.355 | 0.352 | 0.676 |
| R²       | 0.221                        | 0.222        | 0.218                       | 0.18         | 0.18       | 0.197       | 0.207   | 0.207   | 0.25    | 0.205   | 0.205   | 0.25   |

Notes: The dependent variable in models (a) and (b) is the level outcome in the post period. Model (b) includes controls for the pre-Notification period outcome in levels. In model (c) the dependent variable is the linear difference in outcomes relative to pre-Notification period for each officer. All regressions include experimental block fixed effects. Robust standard errors in outcomes allow for arbitrary correlation within officer (320 clusters). The reported control mean is calculated in the pre-notification period.

+ p < .10  * p < .05  ** p < .01  *** p < .001
Table 4: Supervisory Meetings and Other Incident Outcomes (n=1,434)

<table>
<thead>
<tr>
<th>Model</th>
<th>One Week</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>After Notification (a)</td>
<td>(b)</td>
<td>(c)</td>
<td>After Meeting (a)</td>
<td>(b)</td>
<td>(c)</td>
<td>After Notification (a)</td>
<td>(b)</td>
<td>(c)</td>
<td>After Meeting (a)</td>
<td>(b)</td>
</tr>
<tr>
<td>Non-Criminal Infractions</td>
<td></td>
<td>0.011</td>
<td>0.0066</td>
<td>-0.0068</td>
<td>0.021*</td>
<td>0.016</td>
<td>0.0034</td>
<td>0.011+</td>
<td>0.0075</td>
<td>0.0056</td>
<td>0.012+</td>
<td>0.0084</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[0.011]</td>
<td>[0.011]</td>
<td>[0.013]</td>
<td>[0.010]</td>
<td>[0.010]</td>
<td>[0.013]</td>
<td>[0.0063]</td>
<td>[0.0052]</td>
<td>[0.0057]</td>
<td>[0.0066]</td>
<td>[0.0056]</td>
</tr>
<tr>
<td>P(Effect = 0)</td>
<td></td>
<td>0.337</td>
<td>0.531</td>
<td>0.609</td>
<td>0.0428</td>
<td>0.106</td>
<td>0.789</td>
<td>0.0735</td>
<td>0.145</td>
<td>0.330</td>
<td>0.0701</td>
<td>0.133</td>
</tr>
<tr>
<td>Pre-Notification Control Mean</td>
<td></td>
<td>0.224</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>R²</td>
<td>0.24</td>
<td>0.27</td>
<td>0.16</td>
<td>0.33</td>
<td>0.37</td>
<td>0.2</td>
<td>0.45</td>
<td>0.68</td>
<td>0.22</td>
<td>0.48</td>
</tr>
<tr>
<td>Verbal Warnings</td>
<td></td>
<td>-0.003</td>
<td>-0.0025</td>
<td>-0.0006</td>
<td>0.0034</td>
<td>0.0039</td>
<td>0.0058</td>
<td>-0.0035</td>
<td>-0.003</td>
<td>-0.0027</td>
<td>-0.0015</td>
<td>-0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[0.005]</td>
<td>[0.005]</td>
<td>[0.007]</td>
<td>[0.006]</td>
<td>[0.006]</td>
<td>[0.008]</td>
<td>[0.0044]</td>
<td>[0.0028]</td>
<td>[0.0029]</td>
<td>[0.0046]</td>
<td>[0.0032]</td>
</tr>
<tr>
<td>P(Effect = 0)</td>
<td></td>
<td>0.576</td>
<td>0.607</td>
<td>0.935</td>
<td>0.580</td>
<td>0.500</td>
<td>0.453</td>
<td>0.432</td>
<td>0.283</td>
<td>0.347</td>
<td>0.742</td>
<td>0.743</td>
</tr>
<tr>
<td>Pre-Notification Control Mean</td>
<td></td>
<td>0.052</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>R²</td>
<td>0.23</td>
<td>0.28</td>
<td>0.14</td>
<td>0.21</td>
<td>0.26</td>
<td>0.15</td>
<td>0.26</td>
<td>0.63</td>
<td>0.13</td>
<td>0.28</td>
</tr>
<tr>
<td>Assistance Rendered</td>
<td></td>
<td>0.035*</td>
<td>0.031*</td>
<td>0.023</td>
<td>0.04**</td>
<td>0.04**</td>
<td>0.029+</td>
<td>0.0097</td>
<td>0.0076</td>
<td>0.0066</td>
<td>0.0012</td>
<td>-0.0008</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[0.014]</td>
<td>[0.013]</td>
<td>[0.015]</td>
<td>[0.013]</td>
<td>[0.013]</td>
<td>[0.016]</td>
<td>[0.0091]</td>
<td>[0.0073]</td>
<td>[0.0080]</td>
<td>[0.0092]</td>
<td>[0.0073]</td>
</tr>
<tr>
<td>P(Effect = 0)</td>
<td></td>
<td>0.0154</td>
<td>0.0205</td>
<td>0.131</td>
<td>0.002</td>
<td>0.003</td>
<td>0.0594</td>
<td>0.291</td>
<td>0.299</td>
<td>0.414</td>
<td>0.895</td>
<td>0.908</td>
</tr>
<tr>
<td>Pre-Notification Control Mean</td>
<td></td>
<td>0.474</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>R²</td>
<td>0.24</td>
<td>0.31</td>
<td>0.18</td>
<td>0.31</td>
<td>0.35</td>
<td>0.2</td>
<td>0.31</td>
<td>0.57</td>
<td>0.21</td>
<td>0.39</td>
</tr>
</tbody>
</table>

Notes: The dependent variable in models (a) and (b) is the level outcome in the post period. Model (b) includes controls for the pre-Notification period outcome in levels. In model (c) the dependent variable is the linear difference in outcomes relative to pre-Notification period for each officer. All regressions include experimental block fixed effects. Robust standard errors in outcomes allow for arbitrary correlation within officer (320 clusters). The reported control mean is calculated in the pre-notification period.

+ p < .10 * p < .05 ** p < .01 *** p < .001
<table>
<thead>
<tr>
<th></th>
<th>Panel A: Officer Activity</th>
<th>Panel B: Incident Outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>After Notification</td>
<td>After Meeting</td>
</tr>
<tr>
<td></td>
<td>Events</td>
<td>On Views</td>
</tr>
<tr>
<td>Effect of Supervisory Meeting</td>
<td>0.672</td>
<td>-0.00541</td>
</tr>
<tr>
<td></td>
<td>[6.575]</td>
<td>[0.00769]</td>
</tr>
<tr>
<td>Supervisory Meeting x Predicted Risk</td>
<td>-1.09</td>
<td>0.00472</td>
</tr>
<tr>
<td></td>
<td>[5.576]</td>
<td>[0.00825]</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.199</td>
<td>0.281</td>
</tr>
</tbody>
</table>

|                          | Panel B: Incident Outcomes |
|                          | After Notification         | After Meeting             |
|                          | Arrests                   | Force                     | Complaints   | Arrests     | Force       | Complaints  |
| Effect of Supervisory Meeting | -0.00576+                | -0.0701                  | 0.0108       | -0.00754+   | -0.057      | 0.03        |
|                          | [0.00345]                 | [0.0480]                 | [0.0316]     | [0.00403]   | [0.0486]    | [0.0319]    |
| Supervisory Meeting x Predicted Risk | 0.00750*                | -0.0558                  | 0.0296       | 0.00859**   | -0.0281     | 0.0529*     |
|                          | [0.00293]                 | [0.0569]                 | [0.0270]     | [0.00315]   | [0.0777]    | [0.0258]    |
| $R^2$                    | 0.192                     | 0.181                    | 0.230        | 0.184       | 0.208       | 0.172       |

Notes: The dependent variable is the linear difference in outcomes relative to pre-Notification period for each officer. All regressions include experimental block fixed effects, and first order effects of predicted risk. Predicted risk scores are normalized to have a mean of zero and a standard deviation of 1. Robust standard errors in outcomes allow for arbitrary correlation within officer (320 clusters)

$+ p< .10 \quad * p<.05 \quad ** p<.01 \quad *** p<.001$
Table 7: Officer Survey Responses

Panel A: To what extent did this meeting with this supervisor …

<table>
<thead>
<tr>
<th>Question</th>
<th>N</th>
<th>Median</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Make you feel like what you had to say actually mattered to this supervisor?</td>
<td>61</td>
<td>4.0</td>
<td>4.0</td>
<td>1.05</td>
</tr>
<tr>
<td>Make you wish more supervisors in the SPD used this type of feedback approach?</td>
<td>61</td>
<td>3.0</td>
<td>3.07</td>
<td>1.33</td>
</tr>
<tr>
<td>Make you feel more respected than in past supervisory meetings?</td>
<td>56</td>
<td>3.0</td>
<td>2.75</td>
<td>1.46</td>
</tr>
<tr>
<td>Feel like a fair and just system for providing feedback to officers?</td>
<td>56</td>
<td>4.0</td>
<td>3.55</td>
<td>1.13</td>
</tr>
<tr>
<td>Help to stimulate your memory of the event(s)?</td>
<td>65</td>
<td>4.0</td>
<td>3.48</td>
<td>1.17</td>
</tr>
<tr>
<td>Allow you to recall specific details of the event(s)?</td>
<td>65</td>
<td>4.0</td>
<td>3.45</td>
<td>1.20</td>
</tr>
<tr>
<td>Allow you the chance to explain your perspective without feeling judged?</td>
<td>64</td>
<td>4.0</td>
<td>3.94</td>
<td>1.20</td>
</tr>
<tr>
<td>Make you feel you were being unduly criticized?</td>
<td>66</td>
<td>1.0</td>
<td>1.42</td>
<td>1.10</td>
</tr>
<tr>
<td>Help you to understand ways in which you could improve future encounters?</td>
<td>64</td>
<td>3.0</td>
<td>2.77</td>
<td>1.26</td>
</tr>
</tbody>
</table>

Panel B: Open Ended Comments

| Incidents too mundane, low stress, or not relevant | 15  |
| Good/Great idea or tool                            | 7   |
| Need entire squad/all officers on scene            | 4   |
| Current supervisors already do this                | 3   |
| Confidentiality concerns                          | 2   |
| Not beneficial                                    | 2   |
| Neutral                                           | 1   |

Notes: Item response rates ranged from 23.2% to 26.5%
Appendix A: Variables included in CAD incident histories

Based on our conversations with SPD staff about sources of stress and concern for officers, we constructed our history using the following variables:

1. The number of times officers had been dispatched to that place (street segment or census block)
2. The number of 911 calls initiated from that place
3. The number of events in that place involving someone in mental distress
4. The number of times events in that place have involved citizens with firearms
5. The number of times officers have responded to domestic violence incidents in that place
6. The number of times dispatchers have chosen to include officer safety warnings for officers deployed to that place
7. The fraction of events that are correctly described by the dispatcher to the officer
8. The fraction of events that are described by the officer as more urgent, or a higher priority call, than initially described by the dispatcher
9. The number of officers injured on that street segment during the previous month

Since the aim of this model was prediction rather than interpretation of marginal effects, we construct each of these nine history measures at the street segment and census block level over the course of the previous month, but also during the same days of the week and time of day of the specific CAD incident. For example, a CAD incident occurring on Saturday at 3 pm would have four street segment history measures: one measure describing the entire month, one describing the previous four Saturdays, one describing all previous afternoons, and one describing all previous Saturday afternoons. Four block level history measures were created in the same way.
At the request of the SPD, we also included a limited number of officer-specific measures or potential stress, as measured by the department’s Human Resources Department. Here, we paid careful attention to factors that were suspected to be related to an officer’s level of energy, cognitive capacity, and stress level on the job. We also wanted to focus on measures the SPD could respond to, by, for example, offering extended sick leave or limiting overtime, rather than officer specific measures that the SPD has limited control over, such as officer race or gender. We also wanted to exclude measures that were predicted outcomes of our experiment (arrests, use of force, complaints against that officer), as this would introduce path dependence into our randomization process.

10. Whether or not the involved officer had more than the average (median) number of sick days in the previous month

11. Whether or not the involved officer worked more than the average (median) amount of overtime in the previous month

12. Whether or not the involved officer had more than the average (median) number of second jobs during the previous month
Appendix B: Procedural Justice Training Toolkit Excerpt and Sergeant Script for Procedural Justice Supervisory Meeting

Overview of the LEED Engagement

At the beginning of the engagement, the sergeant is asked to follow standard introductory instructions, adapted from Rosenbaum and Lawrence (2011) and Fisher and Geiselman (1992). These consist of:

1. Thanking the officer for coming in to help you test this new supervisory training program: remind them that they were selected randomly, participant identity and comments are confidential and non-attributive, and that this meeting is entirely meant to help examine new supervisory techniques.

2. Transferring control to the officer: tell the officer that you would like to discuss an interaction that they had on (date of flagged encounter) at around (time of flagged encounter), as reported in G.O. # ____. Let the officer choose the starting point for the narrative and give the account at his or her own speed and in his or her own words. Do not interrupt the officer, if at all possible. Listen actively to what he or she has to say. Allow for pauses.

[The following script was provided to sergeants to help guide them through the debrief process].

Debrief Model
Sample Script

Introduction script: Welcome officer and confirm baseline understandings of process and protections.

- Incident # __________ has been randomly selected for a debrief of the circumstances of the incident as perceived and reported by [you] the officer.
- In addition to being a random assignment, officer identity is anonymous and confidential. The department has no access to the identity of the participants in the debrief, and the involved officer will be coded by a code assigned and retained by Cornell University, which again will not be accessible to the department.
- No notes will be taken by the sergeant and lieutenant, and any subsequent documentation of training and equipment needs, tactics, best practices and other valuable insights intended to reduce officer risk in difficult encounters will not include personal identifying information. The watch/operations commander is charged with ensuring that

---

45 In Washington State, procedural justice is frequently referred to as LEED, an acronym for “listen and explain with equity and dignity,” which was developed by the King County Sheriff in 2011 (Rahr, Diaz, and Hawe, 2011).
confidentiality and officer anonymity is maintained at all points of this debriefing process.

At this point, the sergeant should inquire if the officer is aware of any complaint or investigation into the incident being examined, and terminate the debrief if the answer is yes.

- The debrief will take no more than 30 minutes, unless by mutual agreement between [you] the officer and [myself] the sergeant.
- Do you have any questions up to this point?

Sergeants received training on how to answer questions such as: 1) why are we doing it? 2) why all the protections? 3) why is the Lieutenant here? 4) what is LEED? And other questions as they arose regarding confidentiality and anonymity, what constituted a high-risk circumstance, why there was a random and control group, what the union jurisdiction was in this process etc.

- If there are not any questions [or any more questions] I will use some key questions to facilitate your analysis of G.O. #____________.

**Series one script:** Basic incident information, risk assessment and tactics

**Officer orientation:** Describing the incident and thought process in the real time framework of the incident as it unfolded

- How did the incident begin? What were your initial thoughts based on the information you had, or what you initially observed, about the risk factors of this scenario?
- What did you observe about the behavior, mental or emotional state of the subject you encountered? Were there other risk factors observed (drugs, alcohol)?
- Were there external or environmental factors you included in your risk assessment and preliminary tactical thought process? These could include anything which you factored into your size up of the incident [such as] weather conditions, lighting, traffic, presence of hostile (or supportive) bystanders, availability of back-up, confidence in the capabilities of your secondary officers or partner, prior adverse publicity about the location, scenario or involved parties, etc.
- Describe your tactical approach. What was your thought process around how to address the scenario, and the behavior of the subject?
- Did your tactical approach work to address the risks you identified?
• Did your initial risk assessment change? In what way? Were you required to change tactics in response to a change in behavior or risk? What were those changes?
• [Did anything] occur during the encounter, which you would describe as surprising and unpredictable? [If so, what?] How did you respond to these unforeseen events?
• Are there other factors, indicators, risks or other information you consider important in your recollection of the incident?
• Please share your professional assessment of how this incident [was] resolved.

Series two: Post incident assessment and examination of options and needs

Officer orientation: analysis of the incident in retrospect, with an emphasis on identifying ways to mitigate risk and improve department training and support

• In hindsight, do you believe you had the information you needed to size up the situation, identify the risk factors and employ the tactics you needed to address the situation? What, if anything, was missing?
• Do you consider the outcome of the incident to [have been] successful?
• How do you personally and professionally measure success?
• In hindsight, were there things you might have done differently? To put [it] another way, were there options to resolve the incident, which might have been effective?
• Were there options you wish you had? What are they?
• There is a lot of talk in the department around Procedural Justice, otherwise known as “LEED”. What do you know about it? How would you describe it? Based on what you know, how do you assess its utility, relevance or general value?
• How would you size up the options or opportunities to employ LEED tactics in the incident we are discussing?
• Using as your reference point the incident we are debriefing, what can or should the department do to improve your ability to carry out your duties? This is an open question [for example] training, staffing, deployment, equipment, supervision, command, communication, accountability etc. Nothing is off limits.

Note to sergeant: it might be wise to set a time parameter for this question. Remember the 30 minute debrief objective, although this can be exceeded by mutual agreement

• In your professional judgment, what are the greatest risks you confront in your role as patrol officer? And what are the most effective and significant ways the department can reduce those risks?

Series three: Debriefing the brief
**Officer orientation:** candid and constructively critical.

- What questions were missed?
- What information do I [the sergeant] need to know to better meet my role and responsibility to supervise?
- Did this debrief meet your standards and expectations for respectfulness? Relevance? Value?
- How would you change this process? Should this debriefing model be continued? Expanded?
- Any other feedback is welcome.

[END OF SCRIPT]
Appendix C: Distribution Pre-Post Change in Use of Force Incidents

![Histogram showing distribution pre-post change in use of force incidents]