# A Cognitive View of Policing*

Oeindrila Dube[†]
NBER and University of Chicago

Sandy Jo MacArthur[‡]
California Southern University

Anuj K. Shah[§]
University of Chicago

May 31, 2024

**Abstract**

What causes adverse policing outcomes, such as excessive uses of force and unnecessary arrests? Prevailing explanations focus on problematic officers or deficient regulations and oversight. Here, we introduce a new, overlooked perspective. We suggest that the cognitive demands inherent in policing can undermine officer decision-making. Unless officers are prepared for these demands, they may jump to conclusions too quickly without fully considering alternative ways of seeing a situation. This can lead to adverse policing outcomes. To test this perspective, we created a new training that teaches officers to more deliberately consider different ways of interpreting the situations they encounter. We evaluated this training using a randomized controlled trial with 2,070 officers from the Chicago Police Department. In a series of lab assessments, we find that treated officers were significantly more likely to consider a wider range of evidence and develop more explanations for subjects' actions. Critically, we also find that training affected officer performance in the field, leading to reductions in uses of force, discretionary arrests, and arrests of Black civilians. Meanwhile, officer activity levels remained unchanged, and trained officers were less likely to be injured on duty. Our results highlight the value of considering the cognitive aspects of policing and demonstrate the power of using behaviorally informed approaches to improve officer decision-making and policing outcomes.

*This project was supported by the National Collaborative on Gun Violence Research, Motorola Solutions Foundation, and the Griffin Foundation. We are grateful to the Chicago Police Department for their partnership in this work. We are indebted to Daniel Godsel, whose leadership was critical in launching and shaping the Sit-D program. We also thank Jack Benigno, Tom Gaynor, Kristina Knapcik, Chris Pillow, and Trak Silapaduriyan for their extensive guidance in developing and administering the training, and Antoinette Ursitti for essential support in carrying out the evaluation. This work was also made possible by vital support from the University of Chicago Crime Lab, particularly Roseanna Ander, Ashna Arora, Ricardo Avalos, Jordan Bellquist, Anthony Berglund, Hye Chang, Brian Davis, Dylan Fitzpatrick, Hays Golden, John Greer, Katie Larsen, Leah Luben, Sean Malinowski, Peyton Morgan, Ashley Motta, Gonzalo Moromizato, Emma Nechamkin, Khoa Nguyen, Ashley Orosz, Zoe Russek, Greg Stoddard, and Haz Yano. For thoughtful feedback on the study, we thank Bocar Ba, Phil Cook, Felipe Goncalves, Alex Imas, Walter Katz, Jens Ludwig, Sendhil Mullainathan, Jonathan Mummolo, Devin Pope, Bernd Wittenbrink, and participants of seminars and conferences at NBER-Political Economy, NBER-Crime, IOG-BFI, Brown, Cornell, Dartmouth, Princeton, Stanford, UCLA, UCSD, West Point, NCGVR, AL CAPONE and the BFI Economics of Crime and Justice Conference. This project received research approval at the University of Chicago from the Social and Behavioral Sciences IRB, protocol ID IRB18-1234. It is registered with the American Economic Association RCT Registry, RCT ID AEARCTR-0011730. The findings and opinions expressed here are those of the authors and do not necessarily reflect those of the Chicago Police Department.

[†]odube@uchicago.edu

[‡]sandyjo.macarthur@gmail.com

[§]anuj.shah@chicagobooth.edu

# 1 Introduction

Policing practices have increasingly come under public scrutiny, spurring widespread calls for police reform. There is a growing recognition that adverse policing outcomes, such as excessive uses of force and unnecessary arrests, are socially costly for the most heavily policed communities (Ang, 2020; Bor et al., 2018; Weitzer and Tuch, 2004). These adverse events have led to protests and eroded trust in policing across the U.S. (Chen et al., 2021; Williamson et al., 2018; Haseman et al., 2020; Desmond et al., 2016; Jones, 2022; Schuck and Rosenbaum, 2005). They have also proven costly for police departments themselves, resulting in lawsuits and substantial settlements (Alexander et al., 2022; Schwartz, 2016).

There are two common views on the drivers of adverse policing outcomes. First, they might be driven by problem officers—those who are prone to using excessive force or making arbitrary arrests, perhaps ignoring department policies or even allowing explicit or implicit prejudice to shape their actions. Indeed, by some estimates, just 2% of officers are responsible for 50% of instances of misconduct (Walker et al., 2001), and a large literature describes the role of racial bias in policing (Correll et al., 2007; Fryer, 2019; Rozema and Schanzenbach, 2019; Goncalves and Mello, 2021; Hoekstra and Sloan, 2022; Fagan and Campbell, 2020). Second, these adverse outcomes might stem from poor regulations. For instance, department policies can affect whether officers use more forceful tactics (Mummolo, 2018), while a lack of accountability or oversight can open the door to further misconduct (Rivera and Ba, 2022; Rad et al., 2023; Moreno-Medina et al., 2024).

Clearly, both of these views are important for understanding why adverse policing outcomes might arise. Yet, by focusing on individual officers or department-level regulations, these existing views may actually overlook a key aspect of policing itself. Police work often involves making complex decisions in situations that produce stress, trigger many emotions, and require officers to act quickly.[1] These cognitive demands make it more likely that offi-

[1]See Fearon (2019) for a theoretical account of how emotions like fear can lead to adverse outcomes in policing.

cers will act without sufficient deliberation and that their actions will be driven by cognitive biases. In this paper, we explore this overlooked perspective and present a cognitive view of policing.

To appreciate this perspective, consider two scenarios. Imagine there is a 911 call about a man with a gun. When an officer arrives at the scene, he sees a large man in a dark alley who is shouting loudly. If the officer sees a glint of something in the man's hand, he might conclude that the glint is a gun, and he might draw (or even fire) his weapon in response. If it turned out that the glint was not a gun, the officer would have made a mistake. Or, consider an officer who sees a teenager toss a bottle into the street. The officer shouts and tries to initiate a stop, but the teen takes off running. The officer might decide that the teen is guilty of something other than littering, chase after him, and then arrest him for obstructing an officer. But if the teen's greatest offense was littering, this would arguably have been an unnecessary arrest.

Despite the differences in severity, both scenarios produce adverse consequences. We suggest an additional candidate explanation for these outcomes, beyond ill intent or inadequate department policies. Our explanation focuses on features inherent in these types of police encounters. First, officers have to process a lot of information to properly diagnose what is happening. Second, these situations are stressful or otherwise emotionally loaded, and they frequently impose time pressure. These features make policing scenarios cognitively demanding.

As decades of psychological research shows, cognitive demands undermine decision-making (Simon, 1955; Payne et al., 1993; Shah and Oppenheimer, 2008; Kahneman, 2011), leading people to rely on quick, intuitive judgments (i.e., System 1 thinking), rather than more deliberative responses (i.e., System 2 thinking). People may rely too much on their initial assumptions (Nickerson, 1998; Johnson-Laird, 1983) and make judgments based on superficial factors rather than more diagnostic information (Chaiken, 1980; Petty and Cacioppo, 1986). Moreover, people may judge others based on stereotypes (Fiske and Neuberg, 1990)

or without fully considering another person's circumstances (Gilbert et al., 1988). In short, cognitive demands can lead people to narrowly interpret the situations they encounter (Fischhoff et al., 1978; Shaklee and Fischhoff, 1982; Dunning et al., 1990).

This response can be particularly consequential in the policing context. Officers often need to think through multiple *alternative interpretations*, or different explanations for unfolding events. For example, they may need to reconsider critical details which affect the level of force they use, or consider different perspectives on why the subject is behaving in a particular way. In the alley, the officer might initially believe the glint to be a gun, but he also needs to consider the possibility that the glint is just a bottle or a cell phone. In the street stop, the teen might be running because he is scared, not because he is guilty of any other offense. But when facing the cognitive demands of policing, officers may act without considering enough interpretations of the situation. And this can lead to mistakes, negative interactions, and adverse outcomes.

Of course, it is not possible to remove cognitive demands from policing. But it might be possible to improve policing outcomes by training officers to better navigate these cognitive demands. Indeed, field research has shown how behaviorally informed interventions that promote System 2 thinking can effectively reduce crime and violence among youth (Heller et al., 2017) and adults (Blattman et al., 2017; Bhatt et al., 2023), in part by training them to question their automatic assumptions and to be more deliberate in their decision-making.

To test this idea, we developed and evaluated a new training, called Situational Decision-making (Sit-D), which combines a deep understanding of day-to-day policing with insights from behavioral science on how to train people to more deliberately process information and make decisions. The Sit-D training aims to help officers go beyond their initial impression of cognitively demanding situations and develop alternative interpretations.

The training first teaches officers to recognize the kinds of situations that might cause stress and impose cognitive demands. It then teaches officers about specific cognitive biases they may experience in these situations, such as catastrophizing (assuming the worst possible

outcome), personalizing (assuming someone is trying to antagonize them), or engaging in confirmation bias (focusing primarily on evidence that supports their assumptions). Finally, Sit-D teaches strategies to reduce these biases by developing alternative interpretations (e.g., distinguishing between subjective perceptions and objective facts, looking for information that might disprove their assumptions). More generally, throughout the training, officers learn to ask themselves, "What else could I be missing?"

We evaluate the training using a large-scale randomized controlled trial with officers from the Chicago Police Department (CPD)—the second largest police department in the U.S. Our sample comprises 2,070 officers—nearly one-fifth of all sworn personnel in the department. The sample includes all active duty police officers who have been on the job at least two years and who completed a set of mandatory courses.

The evaluation of Sit-D serves two purposes. First, it tests the theory that if officers are unprepared to navigate cognitive demands, this can lead to adverse outcomes (as officers may consider too few alternative views). Second, it serves as a proof-of-concept for how to use training to mitigate adverse outcomes in policing.

The evaluation uses two data sources. First, four months after the training, officers completed an endline assessment that focused on our proposed mechanism—the extent to which officers consider alternative interpretations. The assessment contains a wide array of new measures including survey items, hypothetical vignettes, and simulator exercises.

Across a number of measures, we find that, compared to control officers, Sit-D officers more fully think through alternative interpretations of situations. Sit-D officers consider a wider range of possible motivations behind a person's behavior, they recall more information that goes against their initial assumptions, and they are more likely to update their responses as situations change.

To examine if Sit-D correspondingly reduced adverse outcomes in the field, we analyze CPD's administrative data four months after the training, which aligns with the timing of the endline assessment. Although we cannot measure *excessive* force or *unnecessary* arrests

(as such determinations require in-depth investigations of individual cases), we can measure uses of force more generally and arrests that plausibly could have been avoided. We find that the training leads to reductions in two key adverse outcomes. First, it reduces uses of non-lethal force by 23%.[2] Second, we also examine Sit-D's effects on a pre-specified category of discretionary arrests, which include charges such as disobeying a police officer and disorderly conduct. Many of these arrests likely stem from officers' emotional responses, such as frustration with a subject's behavior.[3] We find that the training leads to a 23% reduction in these discretionary arrests.

Strikingly, we find that Sit-D also mitigates racial disparities in policing. For instance, the overall reduction in discretionary arrests is almost entirely due to the reduction in discretionary arrests of Black subjects. Additional analyses reveal that Sit-D leads to an 11% reduction in overall arrests of Black subjects, without exerting any corresponding effects on arrests of White subjects, or subjects of any other races. This is notable given that the training does not explicitly focus on racial biases or disparities in policing. But it is possible that by making officers more deliberative in general, this could prevent implicit biases from affecting officers' actions (Axt and Lai, 2019).

These reductions may raise questions about whether the training reduces how active officers are or jeopardizes their safety. But we find that there is no reduction in overall officer activity (measured through a pre-specified index of items such as firearm recoveries, drivers' stops, warrants, and citations). In addition, we find that Sit-D reduces the number of days officers take off due to injury on duty in the key four-month period after the training.

To gauge how long the effects of the training last, we also analyze administrative data for our main outcomes 5-8 months and 9-12 months after the training ends. These results sug-

[2]We are not powered to detect effects on lethal force. In our sample of 2,070 officers, there were 20 incidents of lethal force in the year after the training.

[3]These charges are typically for minor offenses, in contexts where the officer could have chosen to resolve the situation differently. As such, they can be viewed as plausibly unnecessary, and previous research has shown that arrests for low-level offenses have little public safety value, defined as affecting the most serious forms of violent crime that drive the overall costs of crime to society (Harcourt and Ludwig, 2006; Chalfin and McCrary, 2018; Chalfin et al., 2022).

gest that the effects diminish over the year, though they do not provide a clear-cut answer to precisely when this happens. For uses of non-lethal force, the treatment effect is statistically insignificant for these additional time periods, but the estimates also have large confidence intervals and are not significantly different from the four-month effect. For discretionary arrests, the 5-8 month treatment effect is statistically significant prior to adjusting for multiple inference, while the 9-12 month estimate is not; though this estimate also does not differ significantly from the earlier period effects. Thus, statistically speaking, we are not able to definitively say exactly when fade-out occurs, though the pattern of results indicates that refresher trainings (which are common in the policing context) will be needed to reinforce the effects over time.

We find that the cost of Sit-D per trained officer is similar to the cost of other trainings from large police departments, but there is no corresponding evidence of the effectiveness of these other trainings. The benefits of Sit-D are more diffuse and harder to value. However, even if we narrowly limit our analysis of benefits to the reduction in officer injuries, we find that the cost of Sit-D per officer trained ($807-$864) is more than offset by the savings from reduced injuries per officer ($1057). Of course the biggest part of the benefits of Sit-D will emerge from reduced uses of force and discretionary arrests, so this simple calculation vastly understates the ratio of benefits to costs to society, but it helps to highlight the cost effectiveness of this training.

Future work will need to address questions about the period over which these effects might be sustained and examine the ideal mechanics of the training (e.g., the intensity and timing, or whether refresher trainings are needed to maintain these effects). But, as a proof-of-concept, our results show that officer behavior is remarkably elastic with respect to this type of training.

Broadly speaking, our findings contribute to the growing literature in behavioral economics documenting various ways in which narrow thinking constrains people's decision-making (Gabaix, 2019). For example, in complex situations, people tend to focus only on

what is in front of them (Enke, 2020); and salient information appears to have outsized influence on outcomes as varied as investment (Bordalo et al., 2013), judicial decision-making (Bordalo et al., 2015) and stereotyping (Bordalo et al., 2016). Here, we consider how mitigating cognitive biases can help broaden people's understanding of different situations.

Critically, we situate these themes in an important context—training police officers to think differently on the job. And, our work makes important contributions toward understanding how to train officers. Related trainings are rarely evaluated rigorously at scale. In fact, we are not aware of any past large-scale RCTs of police training programs that have demonstrated significant reductions in uses of force. Instead, many trainings are widely adopted even though there is little evidence they are effective (or even evidence that they are not effective). For instance, many police departments have some form of de-escalation training despite few rigorous evaluations and mixed results from the evaluations that do exist (see Engel et al. (2020, 2022)).

Meanwhile, there is substantial evidence on the effectiveness of procedural justice training (McLean et al., 2020; Rosenbaum and Lawrence, 2017; Schaefer and Hughes, 2019; Skogan et al., 2015; Owens et al., 2018; Canales et al., 2020; Wood et al., 2020, 2021; Weisburd et al., 2022). However, those trainings have a different substantive focus: They emphasize rules of engagement and prescribe how officers should interact with civilians to engender trust.[4] Sit-D, in contrast, does not prescribe what officers should do, but rather provides general guidance around how to make decisions more deliberately in cognitively demanding situations.

Relatedly, Owens et al. (2018) evaluates a training in which supervisors ask officers to reflect on how they made decisions during recent policing situations. The study finds significant reductions in the likelihood of officers making arrests 6 weeks after the conversations take place—results which are highly encouraging and important. While that training also promotes officer reflection, it differs from Sit-D in that supervisors model principles of pro-

[4]Related work by Banerjee et al. (2021) shows how soft skills training can improve officer communication with victims, as well as victim satisfaction with the police.

cedural justice during the conversations, and encourage officers to adopt these principles in their interactions. Sit-D focuses instead on making officers aware of cognitive biases that might undermine their decision-making, and it aims to mitigate these biases by teaching officers to consider alternative interpretations.

In this way, Sit-D also contributes to the growing literature on behaviorally informed violence-reduction programs aimed at promoting System 2 thinking (Heller et al., 2017; Blattman et al., 2017; Bhatt et al., 2023). While our context is different, it builds on prior research that highlighted the benefits of slower decision-making in moments of conflict. However, that prior work did not specify the key mechanisms for improving deliberation. Our work builds on these results by evaluating a curriculum that is more targeted in its focus on alternative interpretations, and by presenting more direct evidence (from the endline assessment) around this mechanism. These insights could inform a broad range of interventions that aim to improve System 2 thinking. Moreover, we find that this approach to training can mitigate racial disparities in policing outcomes even though it focuses on cognitive biases, not racial biases. This stands in contrast to implicit bias trainings, which are common in police departments even though they appear ineffective (Worden et al., 2020; Lai and Lisnek, 2023). Perhaps, to reduce racial disparities in policing, it may be more effective to disrupt the influence that implicit attitudes have on officers' actions (by making them more deliberative), rather than trying to change those implicit attitudes.

The remainder of the paper is organized as follows. Section 2 provides details on the Sit-D training. Section 3 describes the design and methods used in the study. Section 4 presents the results, and Section 5 provides a discussion of the results and concludes.

# 2 Overview of the Intervention

## 2.1 Training Development, Delivery, and Configuration

Our research team developed the Sit-D curriculum in its entirety. We drew on key concepts from the psychology of decision-making, adapting them to the policing context. We then designed numerous exercises in different formats (detailed below) to make it easier for officers to connect the principles of the training to the issues they face while on duty in the field.

The curriculum design was also iterative. We used a "train-the-trainer" model, instructing 31 CPD trainers on how to deliver the training.[5] During this process, we modified the training based on extensive input from key CPD personnel, including the leadership of the Training Academy. We also modified the training based on events in the city. Notably, after widespread policing protests in Summer 2020, we added more protest scenarios to the curriculum. These steps ensured that the training was relevant and engaging to Chicago police officers.

The training consisted of four sessions that were each four hours (i.e., 16 hours total). Each session targeted having 16 officers and four trainers. This ratio was important for managing the different components of each session and facilitating discussion. Typically, there were several weeks in between each session. This allowed officers to start using lessons from Sit-D while in the field and to begin subsequent sessions by debriefing how they had applied the training. Sessions consisted of a mix of classroom instruction (which included lecture and interactive activities) and scenario-based exercises. The first two sessions had more classroom instruction, while the final two sessions were entirely scenario-based exercises. Officers in the training had to take the first two sessions (which were foundational) before they could move to the final two sessions.

[5] Trainers were assigned to classes by CPD's training division. Trainees were not cohorted, so they took different Sit-D sessions with different officers, and were exposed to different trainers across these sessions. We also do not have information on which trainers taught which trainees in each of the classes, so we are unable to test whether some trainers were more effective than others.

## 2.2 Principles and Activities in the Situational Decision-Making Training

Sit-D's curriculum is based on a core lesson: the importance of developing multiple perspectives on any given situation. Officers are taught that to respond effectively to ambiguous situations, it is critical to go beyond one's first impression and develop additional possible explanations for what is occurring. In this section, we briefly describe the curriculum's main framework, along with descriptions of a few exercises from the training (see Table A1 for a fuller list of sample activities.)

The curriculum is organized around a five-step "Thinking Tactic Model." The first two steps of this framework focus on helping officers recognize and regulate their emotional and physiological responses to policing situations, as these can make it more difficult to think systematically (Bodenhausen et al., 1994; Lerner et al., 2015; Kassam et al., 2009). For instance, in one of the first exercises of the training, officers discuss situations in which they felt civilians showed "contempt of cop." These can often be fairly mundane things, like refusing to show ID or talking back to an officer. This discussion highlights for officers how common it is for policing situations to trigger emotions that might interfere with deliberative thinking. Officers also get extensive practice with various breathing exercises (many of which are done while listening to difficult radio calls) to help regulate their responses to these situations.

The remaining three steps of the Thinking Tactic Model encourage officers to consider alternative interpretations of (and responses to) situations. Instead of focusing singularly on the same thought, they are encouraged to come up with more than one possibility for what they are seeing. And instead of assuming their first impression is correct, officers are told to look at the situation through different perspectives. Then, officers are taught to think through more than one way of responding to the situation. Finally, officers are instructed to think through the consequences of each possible response.

As part of these steps, officers learn about various "cognitive biases" or "thinking traps,"

which are mental shortcuts that might constrain their perspective on a situation. These thinking traps were adapted from Cognitive Behavioral Therapy (where they might be referred to as "cognitive distortions"), and the psychology of judgment and decision-making (where they might be referred to as "heuristics and biases"). Specifically, officers are taught about catastrophizing (assuming the worst possible outcome will occur), minimizing (downplaying potential risks), personalization (assuming others' actions are meant to antagonize oneself), confirmation trap (focusing on information that supports one's assumptions), overgeneralization (basing interpretations too heavily on salient past experiences), all-or-none thinking (thinking in absolutes and ignoring nuances), and anchoring (failing to update one's impression as the situation changes).

Officers discuss situations in which they have found themselves experiencing these thinking traps, as well as how they can notice themselves falling into those traps in the field. They are also taught simple questions to ask themselves to mitigate the thinking traps. For example, a common tactic emphasized here is the "camera view," in which officers are asked to note details of an interaction without imparting any judgment or subjective interpretation. They are reminded that a camera cannot "see" ill intent or disrespect in a subject, it can only see a sequence of actions that subjects undertake. Exercises like this help officers distinguish between their subjective impressions and objective facts, prompting them to explore other ways to interpret situations and subjects' actions.

Moreover, in many exercises officers watch and discuss videos of ambiguous policing situations. To encourage developing alternative interpretations, officers first come up with their own explanation privately, then they debate the interpretations as a group. This highlights how even officers with the same training might see situations differently, and thus there is value in going past one's first impression.

Beyond these classroom exercises, officers practice these principles over the course of approximately 12 Force Option Simulator (FOS) exercises, which we again selected because there are many different ways to interpret the situations as they unfold. During FOS ex-

ercises, officers navigate scenarios by interacting with life-sized subjects projected onto a screen. They can speak to the subjects, whose responses are controlled by a trainer operating the FOS machine. Specifically, the trainer operator can branch the scenarios in different ways, with the subject posing a direct threat in some branches, but not others. In this way, the scenarios typically contain the possibility of potential threat. Officers also have retrofitted equipment (TASERs, firearms, and pepper spray) that they can use during the scenarios.

Importantly, officers actively debrief each exercise with their trainers and other officers in the session. The debriefs are designed to help officers see how cognitive biases may have affected their decision-making and how they can avert these distortions. For example, officers are asked a series of questions that push them to articulate their interpretation of the situation, the evidence for their interpretation, and the reasoning behind their decisions and actions. Critically, they are also asked about other possible interpretations and actions they may or may not have considered, as well as features of the scene they did not mention. These active discussions are intended to surface details that they might not have noticed and interpretations they may not have considered. The discussions also help officers recognize how the force options they employ are tied to their interpretation, and reinforce that officers have multiple force options at their disposal.

Note that Sit-D is not unique in its use of simulator training. Since 2014, all CPD recruits have received simulator training while in the Academy. However, Sit-D differs in its active approach to the debriefs, which are tied to discussions of cognitive biases. Finally, one might wonder whether the sample of scenarios could have shifted officers' priors about the dangers they face in the field. For example, if the training disproportionately sampled from non-threatening scenarios, officers might assume they face fewer risks, and they may use less force or make fewer arrests in the field. However, the majority of scenarios in the curriculum involved some level of threat, so it is unlikely that the sample of scenarios led officers to see fewer risks in the field. More generally, by emphasizing alternative interpretations, the

training conveys the idea that every situation could potentially evolve to contain a threat, or to be innocuous. In that regard, the training can be conceptualized as helping officers make use of their time to draw in more information, consider alternative possibilities, and update their priors about the situations they encounter.

Anonymous course evaluations (administered by CPD to 942 trained officers) suggest that the training was engaging and well taught. For example, 92% of the officers reported that they found the training either useful or very useful; while 83% reported they either liked the training or liked it a lot.

# 3 Design and Methods

To assess the causal effect of the Sit-D training, we implemented a randomized controlled trial with CPD. Below, we provide an overview of our sampling procedure, detail our data collection, verify the integrity of the experimental design, and specify our empirical strategy.

## 3.1 Sampling

**Officers in the Sample.** Our sample comprises CPD officers on active duty[6] who completed these prerequisite courses: Law Enforcement Medical and Rescue Training (LEMART) and three Procedural Justice courses.[7] Specifically, the sample includes 2,070 active-duty police officers who have been on the job for two or more years, including those who work in one of 22 police districts in Chicago, as well as those who work in more specialized units, such as gang units, tactical teams, and area saturation teams.[8] We refer to districts and specialized units as the units of assignment.

[6]At the time of randomization, we excluded any officers on desk assignments and any officers who would be on furlough during the training.

[7]These courses had to be completed by everyone at CPD for the department to meet requirements under the state consent decree. As a result, the Academy favored using these as prerequisites for Sit-D to ensure that Sit-D participation would not delay their completion.

[8]Police recruits undergoing training at the Academy and probationary police officers, who are out of the academy less than a year, are not a part of our sample.

**Stratification and Randomization.** The units of assignment typically have four shifts (also known as watches). We stratified the randomization by unit x watch, which resulted in 92 strata. We used random assignment to select approximately half the officers in each stratum for the training group, while the other half served as the control group. In total, 1,059 officers were assigned to the Sit-D training.

CPD asked us to stratify using this procedure since removing all officers in a given unit-watch from duty for training purposes would potentially jeopardize public safety. Given this approach, it is possible that control officers may influence the outcomes of treatment officers (and vice versa) within a given stratum. To the extent that these spillovers occur, they would lead us to understate the true impact of the Sit-D training. Relatedly, we also considered randomizing partners into treatment status. However, this was not feasible since many CPD officers do not work with regular partners.

Based on the conditions of the consent decree, all CPD personnel were required to complete 32 hours of in-service training in 2020. Sit-D counted toward meeting these 32 hours. Since this was a large increase relative to the previous year (when they had to complete 24 hours), it was challenging for CPD to coordinate this effort institutionally, and officers were just barely able to complete their required hours.[9]

Control officers in our sample do not take one specific counterfactual training, but rather a varied combination of 119 different existing trainings. These range in topic from ongoing policing activities (such as traffic direction and control), to learning new technologies (e.g. the ShotSpotter gunshot detection device), to interacting with specialized populations (e.g.,through the National Association for School Resource Officers training), to other types of scenario training (e.g. for Active Shooter Threats). No particular training shows a concentration, indicating the extent to which officers take varied combinations of these courses. Moreover, the timing of these classes was preset institutionally by CPD, and did not change on account of Sit-D.

[9]Personal communication with the former Deputy Chief of the Training division, 2/14/2021.

Since control officers are taking a varied combination of other classes, this makes it less likely that our estimates reflect the negative effects of all these other trainings, rather than the positive effects of Sit-D. In addition, we measure and present evidence on mechanisms that are specific to Sit-D, which further suggests that the observed changes in officer behavior stem from shifts that reflect key aspects of the Sit-D curriculum.

**Timeline.** We conducted the randomization at the end of February 2020. Sit-D training started in March 2020 but had to be paused after two weeks due to the COVID-19 pandemic. It resumed again in September 2020 when CPD re-started its training activities. Sit-D classes continued until February 2021, though 75% of the treatment group had completed the training by December 2020. Figure 1 presents a consort diagram displaying the timeline, and Figure A1 shows the completion rates for each session over time, along with a summary of concepts introduced in each session.

## 3.2 Data

We use two sources of data for the evaluation. We designed an endline assessment, which was administered over March-July 2021—about four months after treatment officers had completed their last session. We also use CPD's administrative data to track outcomes in the field over this same four-month interval, which constitutes the key evaluation period.[10] We wrote two pre-analysis plans (PAPs) which specified the outcomes we would be analyzing from each data source.[11]

[10] We also use additional administrative data to track outcomes over additional periods 8 to 12 months after the training, through February 2022. This is feasible as fidelity to treatment assignment was maintained until March 2022. But starting then, the control group was potentially exposed to Sit-D, since CPD introduced a new Use of Force training for all officers, which incorporated some cognitive biases (e.g., Anchoring and Personalization) from Sit-D into its curriculum.

[11] The PAP for the endline assessment tool can be found here. The PAP for the administrative data can be found here.

### 3.2.1 Endline Assessment Tool

CPD personnel administered the endline assesment at the training academy. First, officers completed a computer-based survey. Second, officers completed scenario-based exercises in a Force Options Simulator (FOS). Out of 2,070 officers, 1,696 officers completed the endline assessments,[12] and 98% of these assessments were completed in-person at the Academy.[13]

In the main text, we focus our discussion on the sections of the endline assessment that are most pertinent to how the training affects (a) officers' consideration of alternative interpretations and (b) officers' behavior in the FOS exercises. We describe these sections briefly below. For more details on the procedures, see Appendix A.1, which also describes other sections of the endline assessment (such as questions about which concepts officers recall from the training and self-report items on what strategies officers use to regulate stress and emotions).

**Considering Alternative Interpretations.** Three tasks measured the extent to which officers consider alternative interpretations. First, the "Driver's Actions Task" focused on the interpretations officers listed for an ambiguous scene. Officers watched a video clip in which police stopped a driver who immediately jumps out of his car.[14] Officers wrote down as many interpretations of the driver's actions as they could think of. Responses were coded into three categories: (1) The driver needs assistance, (2) Enforcement action is required against the driver, (3) A miscellaneous "other" category. We expected that Sit-D trained officers would offer more varied alternative interpretations (i.e., explanations from more than one category).

Second, the "Pictures Task" focused on the information officers use when assessing ambiguous situations. Officers viewed photos of ambiguous situations, where it was unclear if

[12] As we discuss in Section 3.3 below, this sample is balanced across treatment and control.

[13] 45 officers completed the surveys online in response to an email that CPD sent with a link to the survey, a step that was taken to maximize participation. These 45 officers did not complete the simulation exercises, which had to be done in person.

[14] The video can be found here.

a person in the photos was committing a crime. Officers selected either a criminal or non-criminal interpretation of the person's actions, and they wrote down which features they observed that supported the interpretation they selected ("Confirming features") as well as the interpretation they did not select ("Alternative features"). We expected that Sit-D trained officers would be better than control officers at recalling alternative features because they would have more fully considered each alternative interpretation.

Officers completed two versions of this task. In the 3-second version, officers viewed each photo for three seconds. In the officer-timed version, officers controlled how long they viewed the photos. For this latter version, we recorded their viewing time ("Processing Time Index"). In both versions, we recorded how long officers took to decide on their interpretation ("Decision Time Index").

The third task assessed how officers update their responses to situations in which they might use force. Officers watched brief videos of scenarios, after which they indicated how threatened they would feel if they were in that situation, how the civilian would be categorized according to CPD's Use of Force Policy, and what level of force would be authorized for responding to the civilian. Officers also listed different courses of action they would take (as many as they could think of). These responses were coded as appropriate if they matched Sit-D and Use of Force trainers' responses; otherwise, they were coded as inappropriate.

Importantly, one of the videos was a two-part video. The video paused partway through, and officers were prompted to respond to the questions listed in the previous paragraph. The video then continued and officers responded to the same questions again after it concluded. This was done to assess the degree to which officers update their responses based on how a situation changes. We expected that Sit-D trained officers would update their responses to a greater degree and list more appropriate responses.

**Performance in the Simulators.** In the other main component of the endline assessment, officers completed three FOS exercises. Since Sit-D trained officers had gained more

experience with these simulators, it is possible that we would observe practice effects on this task. However, practice effects would not be a concern for the other tasks described above since they were novel for both treatment and control officers. Officers also did not debrief these FOS scenarios, to ensure that the assessment did not inadvertently act as a training for the control group.

CPD personnel observed and coded whether officers: discharged any weapons (and how many shots were fired if they discharged their gun); communicated with the person; gave verbal direction or issued verbal commands; radioed dispatch; froze during the scenario; knelt to make themselves a smaller target; or moved to cover and concealment. To avoid bias, coding was completed by instructors who taught CPD's Use of Force curriculum (and who were not aware of which officers were in the Sit-D training and control groups). We also measured the extent to which officers shot at those who pose direct threats in the scenarios. We expected that Sit-D-trained officers would be more communicative in the scenarios and that their decisions to shoot would be more sensitive to the threat posed by subjects.

### 3.2.2 CPD's Administrative Data

To assess effects on field outcomes, we use different types of administrative data from CPD.

**Uses of Force.** We use data from Tactical Response Reports (TRRs) to measure uses of force. TRRs provide comprehensive information on force incidents since they must be filled out every time a subject resists, threatens, or physically attacks an officer; or is injured by an officer (Chicago Police Department, 2021).

In the post-training data we analyze, uses of force are divided into three categories: Level 3 comprises lethal uses of force (e.g., police shootings); while Levels 1 and 2 comprise non-lethal uses of force, ranging from use of wristlocks to TASER and OC spray (see Appendix A.2 for more detail).[15] We distinguish between lethal and non-lethal levels of force (per our

[15]Prior to 2020, CPD used a 4-point grouping which does not map cleanly onto the newer 3-point categorization. We use measures from the earlier 4-point grouping for balance statistics. We also create an

PAP) since there are only 8 lethal force incidents in our focal sample period four months after the training, and we are not powered to detect changes in this outcome. We therefore focus on non-lethal uses of force, and measure the number of such incidents associated with each officer in a given month. There are 274 such incidents in the focal period (with 68 incidents occurring on average, per month). We pre-specified examining not just non-lethal force (levels 1 and 2) but also higher levels of force (levels 2 and 3) as well as all levels of force together (levels 1, 2 and 3). We analyze these two latter measures in Appendix B.2.

The TRRs contain other information on subject injury and tactics, which we also analyze in Appendix B.2. These include: Officer recorded injuries, subject allegations of injuries, measures of hospitalization, and an index of officer reliance on force tactics (versus other types of tactics) in use of force incidents. We describe the measurement challenges inherent in these variables in Appendix A.2. Our PAP categorized injuries from TRRs as primary outcomes, but we present them in the appendix given these measurement issues.

**Arrests.** We also draw on CPD's arrest data to examine various types of arrests. As in Rivera and Ba (2022), Ba et al. (2022) and Lum and Nagin (2017), we do not take arrests to be a measure of productivity. In fact, we suggest that some arrests may be unproductive in that the officer could have taken another course of action to resolve the situation effectively (i.e., these arrests are discretionary and plausibly unnecessary). We pre-defined one subset of such discretionary arrests that we hypothesized Sit-D would be most likely to reduce. Namely, these are arrests that occur in situations where the officer may be responding out of irritation or frustration (for example, based on their perception that a subject is being disobedient or disrespectful). These include charges such as obstructing an officer, resisting an officer, disobeying an officer, and various types of disorderly conduct (see Table A2 for the complete list of statutes included in this measure). We hypothesized that the training would reduce this subset of arrests since Sit-D helps officers identify situations in which they are

alignment across the old and new classifications to implement a difference-in-differences analysis (described in Appendix B.2). However, this alignment is noisy and we cannot verify its complete accuracy given the mapping challenge.

likely to personalize situations or perceive "contempt of cop." Moreover, it teaches officers strategies to move past these initial perceptions by considering alternative interpretations of and motivations behind subjects' actions. As such, these types of discretionary arrests are one of our main measures of adverse policing outcomes.

There are 265 discretionary arrests in our sample over the 4 month post-training period. While we defined one particular subset of arrests that are both discretionary and highly relevant to our training (covering a small 1.3% of all arrests), it is not meant to comprehensively span all classes of discretionary arrests, and there may be other subsets of arrests that are discretionary, unnecessary, low-value, or otherwise unproductive. For example, Rivera and Ba (2022) uses a broader classification, which we use to check the robustness of our findings (see Section 4.2).

Our data on arrests also include the race of the arrestee. In our PAP, we did not specify examining separate effects by race as a primary outcome, but instead specified examining race of subject as a dimension of heterogeneity. This was in the interest of pooling observations for power, and because the training did not focus explicitly on racial bias. In examining heterogeneity, we focus on Black subjects and subjects of other races in the main paper, while further disaggregating other races into Hispanic, White, and a miscellaneous "other race" category (which includes Asian/Pacific Islanders and Native Americans) in the appendix.

Importantly, the CPD data we use attribute both arrests and uses of force to all officers involved in an incident, regardless of whether they are designated as the primary, secondary, or assisting officer. This limits the scope for potential manipulation and the possibility that treated officers might disproportionately ask control officers to be designated the primary officer, with the aim of getting incidents assigned to the records of these other officers.

**Other outcomes.** To gauge effects on officer activities more generally, we turn to administrative data from the Performance Recognition System (PRS), and we use it to build an index of officer activity which includes: warrants; recovered vehicles; recovered guns; traffic

stops; driver stops; Investigatory Stop Reports (ISRs)[16]; Administrative Notices of Ordinance Violation (ANOVs); citations; curfew violations; CTA checks; parking citations; and all other arrests that are not a part of our pre-specified category of discretionary arrests. To measure effects on officer injuries, we use daily attendance data, which provides information on days off due to injury on duty (IOD). Note that the index of officer activities is listed as a secondary outcome in our PAP, and officer injuries were not included in our PAP. However, we present these in our main table examining key administrative outcomes because officer activity and injuries are important for contextualizing and interpreting the reductions in uses of force and discretionary arrests.

Based on the theory that we outline in the introduction, we expected the largest effects from Sit-D to be seen for outcomes that arise in ambiguous interactions with civilians where officers have meaningful discretion over how they respond. In such situations, we expected that officers might intuitively favor more severe enforcement responses, but deliberation could lead to alternative ways of interpreting and resolving the situation. Given that uses of non-lethal force and discretionary arrests might often result from situations that have these characteristics, we posited that Sit-D would be most likely to lead to reductions in these outcomes. This does not preclude the possibility that Sit-D would lead to reductions in other actions, but rather indicates where we felt our theory made the clearest predictions.

In the appendix tables, we also examine two additional outcomes that are downstream responses to officers' actions: complaints levied against officers, and awards and commendations given to officers. We discuss measurement issues in the complaints data and challenges to interpreting awards as a measure of performance in Appendix A.2. We initially included complaints as part of our primary outcomes in our PAP, but given these issues, we discuss these outcomes in the appendix.

In the tables below we present all (non-index) outcomes in units of per 1,000 officers per month, with the exceptions of days off for injuries (which are presented per officer per

[16]Our PAP notes that we would include "ISRs/Contact Cards", but CPD replaced Contact Cards with ISRs in 2016, prior to our sample period.

month).

**Families of Outcomes.** Besides grouping together closely-related outcomes into mean effect indices, we also group conceptually related indices and outcomes together into broad families, which we use when adjusting for multiple hypothesis testing. For the endline assesssment, these families include outcomes described above in the main text as well as in the appendix (see Appendix A.1). We create a Knowledge Family, consisting of both the Knowledge of Sit-D Concepts Index and questions related to the knowledge of Use of Force Policy. We also create a Navigating Cognitively Demanding Situations Family, which includes measures of how officers first approach these situations (the Coping With Stress, Emotion Regulation, and Confidence indices), measures of how officers think through alternative interpretations (from the Driver's Action and Pictures Tasks, as well as the use of force videos), and measures of thinking traps that can emerge in these situations (the Personalization Index). Finally, we create an Officer Performance in the FOS Family, which comprises all outcomes from the simulators.

From our main administrative measures, we create an Adverse Policing Outcomes Family, comprising uses of non-lethal force and discretionary arrests, and an Officer Activities And Injuries Family, comprising officer injuries and the index of officer activities. From the additional data, we create an Auxiliary TRR Family, comprising all secondary outcomes from TRRs including injuries and tactics used in these incidents; and a Downstream Outcomes From Officers' Actions Family, comprising commendations and awards and complaints outcomes. Table A3 details the specific indicators in each of these families.

## 3.3 Integrity of the Experiment

**Balance.** Table B1 presents balance across key covariates. In Panel A, we examine key officer characteristics (age, gender, experience, and race). In Panel B, we examine baseline outcomes from the administrative data, including all key variables that we analyze at endline,

for the two years preceding randomization. The table shows balance across these covariates. It also presents a F-test which examines whether the covariates together are jointly significant in predicting the Sit-D treatment indicator. The p-value from this F-test is .39, so we cannot reject the null hypothesis that the covariates together are jointly insignificant. Table B2 verifies that balance is maintained in the subset of 1,696 officers who also completed the endline assessment. Thus, imperfect rates of assessment completion do not affect the integrity of the experiment.

**Attendance and Attrition.** CPD made Sit-D mandatory for those assigned to take the course, which resulted in high rates of training completion (exceeding 86%). In Appendix B.2 we discuss reasons why compliance may not have been 100% (e.g., officers might retire or go on medical leave). Relatedly, in Table B3 we show that attrition did not occur disproportionately out of either the training or control groups.

## 3.4 Empirical Strategy

To gauge the causal effect of the Sit-D training, we estimate Intent to Treat (ITT) effects. To examine outcomes from the endline assessment, which was administered four months after the training, we estimate the following specification:

$$y_{os} = \alpha_s + \beta SitD_o + X_o\delta + \varepsilon_{os} \tag{1}$$

where $y_{os}$ is the outcome for officer $o$ in stratum $s$; $X_o$ is a vector of baseline officer characteristics (discussed below); $SitD_o$ is the treatment indicator, which equals one for officers randomly assigned to the training group, and zero for officers assigned to the control group; and $\alpha_s$ denote stratum (unit x watch) fixed effects. As detailed in Section 3.1 units are either one of 22 police districts or a more specialized unit, while watches correspond to one of four start times. Thus the inclusion of these geography-based stratum fixed effects means that we compare officers in treatment and control who are working in similar environments,

which give rise to similar policing tasks. In addition, Table B4 shows that relative to control, officers in treatment do not switch more to a different unit or unit x watch, from the one in which they were working at the time of randomization.[17] This provides further verification that treated officers do not differentially change their policing environment over the duration of the experiment.

To examine effects on outcomes that are conceptually related to one another, we construct mean effect indices using the approach of Kling et al. (2007). To create an index of $K$ outcomes, we first reverse outcomes where necessary such that a higher (or lower) value consistently indicates better outcomes. We then compute $\widetilde{y}_o = \frac{1}{K}\sum^{K}\left(\frac{y_{ok}-\mu_{0k}}{\sigma_{0k}}\right)$, where $\mu_{0k}$ and $\sigma_{0k}$ are the estimated control-group mean and standard deviation for outcome $k$ in family $K$. Our estimates for these indices thus represent standard deviation changes relative to the control group. Following Kling et al. (2007), when $y_{ok}$ is missing, but another sub-component of the index is measured, we impute the mean from the same treatment arm.

We also examine field outcomes from the administrative data for four months after the training, when the endline assessments were also administered. To examine field outcomes over this key period, we estimate:

$$y_{ost} = \alpha_s + \beta SitD_o + X_o\delta + \gamma_t + \varepsilon_{ost} \tag{2}$$

where $y_{ost}$ is the outcome for officer $o$ in stratum $s$ and month $t$; and $\gamma_t$ are month fixed effects, which account for potential seasonality in policing outcomes. Recall that $SitD_o$ denotes if an officer has been assigned to treatment—i.e., randomization occurs at the level of the officer. In this specification, four monthly post-training observations are included for each officer. Therefore, standard errors are clustered on officer. In the appendix we also check the sensitivity of our results to a hypothetical alternate key evaluation period of

[17] This table presents three different measures of switching, which capture whether the officer was working in a different location in any of the twelve months after the training, all twelve months after the training, or the majority of this post-training period. The top panel measures switching away from a unit x watch and the second from just the unit. The coefficient on the treatment indicator is insignificant and small across all six specifications.

three months, which includes three monthly post-training observations per officer officer; our results are insensitive to this configuration.

Utilizing CPD's administrative data requires us to demarcate the start of the post-treatment period for treatment and control officers. For treated officers, the post-training period starts after they complete their last class (because this is when they are meaningfully trained), which occurs on different dates. We randomly assign control officers to one of these potential training completion dates. This ensures that we have an even number of treated and control officers in each post-training month. While this was the approach pre-specified in our PAP, we verify that our results are robust to using an alternate approach where we instead assign each control officer to all the post-training periods represented among treated officers in their stratum (see Appendix B.2 for greater detail.)

To gauge the period over which effects are sustained, we additionally examine effects eight and twelve months after the training. We do so by pooling all twelve months of post-training data and estimating:

$$y_{osnt} = \alpha_s + \sum_{n=1}^{N}[\theta_n P_{ont} + \beta_n(SitD_o \times P_{ont})] + X_o\delta + \gamma_t + \varepsilon_{osnt} \tag{3}$$

where $P_{ont}$ are period indicators for months 1-4 after the training, months 5-8 after the training and months 9-12 after the training; and $\beta_n$ is the treatment effect in each of these periods. In these specifications twelve monthly post-training observations are included for each officer.

In estimating equations (1)-(3) we include additional covariates, which serve to improve the precision of the experimental estimates. We focus on a control set comprising key officer level characteristics (namely, years of experience, race, and gender) and baseline values of all our main administrative outcomes that are measured in the same way across baseline and endline: discretionary arrests, the index of officer activities, and officer injuries at baseline.[18]

[18] As discussed in Appendix A.2, CPD changed how it measured uses of force between baseline and endline which makes them non-comparable across the two periods.

We have baseline data for a period of two years prior to randomization.

This approach has the advantage that we apply a common set of controls across all estimates in the paper. However, in the appendix, we show that the results are robust to two other specifications—one of which employs the Double LASSO selection technique of Belloni et al. (2013) to select controls, and a second which includes no additional controls.

In addition to conventional standard errors and p-values, we also report q-values that control for the proportion of incorrectly rejected null hypotheses (Benjamini et al., 2006; Anderson, 2008). Specifically, we control for the false discovery rate (FDR) within the period of analysis, across outcomes that are conceptually related to one another under broad families (summarized in Table A3).

# 4 Results

First, we present results from the endline assessment to examine how the Sit-D training affected officers' thought processes. Then, we discuss Sit-D's effects on field outcomes.

## 4.1 Endline Assessment Outcomes

To measure the impact of Sit-D training on assessment outcomes, we estimate equation (1). Here, we focus on the parts of the endline assessment most relevant to how officers consider alternative interpretations and how they navigate scenarios in simulator exercises. For additional endline results, such as concepts that officers recall from the training and self-regulation strategies officers report using, see Appendix B.1.

**Considering Alternative Interpretations.** There are three main tasks in the endline assessment that measure the extent to which officers consider alternative interpretations. The results for these tasks are shown in Table 1. Each panel in the table corresponds to a different task, and each row corresponds to a different outcome.

The top panel shows the results from the Driver's Actions Task. Although Sit-D officers did not generate more total explanations (first row of the panel), they did generate more *varied* explanations of the situation—trained officers were more likely to offer more than one category of explanation (second row). In addition, this effect appears to be driven by the fact that Sit-D officers were more likely to say that the subject might need assistance (third row). In contrast, Sit-D and control officers do not differ in how likely they were to list enforcement-related or other explanations for the subject's actions (bottom two rows).

The middle panel shows the results from the Pictures Task. We find that Sit-D officers were better at recalling and listing information that supported an interpretation of a situation that differed from the interpretation they ultimately chose (first row of the panel). Meanwhile, we do not see significant treatment effects in officers' recall of information that supported their chosen interpretation (second row). Thus, Sit-D officers are better at taking in "disconfirming" evidence, while still recalling the same amount of information that supports their chosen conclusion. This suggests that Sit-D increases the scope of information that officers are taking in.[19] We additionally test whether Sit-D affects how likely officers are to conclude that someone is committing an offense. We find that Sit-D officers were less likely to attribute criminality to subjects' actions in the photos (third row). Thus Sit-D not only affects the information officers take in but also how they integrate that information to come up with an explanation for a situation.[20]

Based on prior work (e.g., Heller et al. (2017); Imas et al. (2022); Brownback et al. (2023)), we expected Sit-D officers to take longer to process scenes and decide on their interpretations. However, we do not find evidence along these lines. Recall that in the Pictures Task officers first view the scenes and then see the two possible interpretations (at which point the timer records how long they take to choose between these possibilities). Treatment and control

[19]Note that due to occasional computer errors in this segment of the endline, there are slightly fewer observations for these measures compared to other endline measures.

[20]These results do not necessarily mean that Sit-D officers will always attribute less criminality to a subjects' actions. Rather, they suggest that Sit-D officers are indeed using the additional information that they notice to guide their assessments. With a different set of photos, Sit-D could have have led to more criminal interpretations.

officers spend the same amount of time viewing the photos—this time is fixed in the 3-second task and treatment officers do not spend more time in the officer-timed task (as shown in the fifth row). Yet, Sit-D officers notice more alternative features of the photos (as the second row indicates) and decide on their interpretations significantly *faster* (fourth row).

One possible interpretation of this finding is that Sit-D may make officers more efficient both in how they process information and decide on an interpretation. For example, trained officers may already have considered different possible interpretations when viewing the scene (as responses from the Driver's Action Task might suggest), and they might therefore spend less time thinking through these possibilities when making a decision. Although this result differs from our predictions, it underscores a point of emphasis in the training: Officers are not told to slow down, but rather to make the most of the time they have.

The bottom panel of Table 1 shows the results from the task involving use of force videos. The top row shows how officers changed their responses to the two-part video in which a subject fires their gun at another person (in part 1), but then drops their weapon and puts their hands up (in part 2). The significant negative coefficient indicates that, upon observing the subject drop their weapon and put their hands up, Sit-D officers lowered their perceived threat, their categorization of the subject, and their chosen force option to a greater degree than did control officers. This suggests that Sit-D officers did not just remain tied to their first interpretation or ignore additional evidence. Rather, they updated their interpretation as the situation changed. Moreover, Sit-D officers listed more appropriate ways of responding to these use of force scenarios (by providing more responses that matched what trainers had listed as appropriate actions). Meanwhile, there were no differences in how many inappropriate actions were listed (bottom two rows of the panel). Thus, not only do Sit-D officers come up with different interpretations of a situation, but they also think of more appropriate ways to respond to it.

Overall, these results provide strong evidence for our proposed mechanism. Sit-D leads officers to come up with more varied interpretations of the cognitively demanding situations

they encounter. The training increases the extent to which officers take in disconfirming information. It improves the extent to which officers update their responses to dynamic situations, and it also enables them to come up with more appropriate responses to situations.

Since trained officers consider alternative interpretations to a greater degree, this may raise concerns that the training leads officers to second-guess themselves, undermining their self-confidence. However, as shown in Table B8, Sit-D officers feel greater confidence in handling their duties, suggesting this is not the case.

In the next section, we consider whether Sit-D also changes officers' behavior when navigating scenarios in the simulator (FOS) exercises.

**Performance in the Simulators.** Table 2 shows the results from the FOS exercises. The top panel of the table shows that Sit-D officers were more communicative and active on a number of dimensions that might help officers respond effectively to situations without necessarily using force. For example they were more likely to give verbal direction to the person with whom they were interacting.

In the bottom panel, we examine whether Sit-D officers become better calibrated in their decisions to shoot—i.e., choosing to shoot more often specifically when faced with a deadly threat. To assess this, we pool the data from the different scenarios, and interact Sit-D with an indicator of whether the subject presents a direct and deadly threat in the scenario (see Appendix A.1 for a description of the scenarios, including which subjects pose a direct threat).

In Table 2, the significant positive coefficient on the interaction term of Direct Threat x Sit-D indicates that trained officers were more likely to fire on those who posed a direct threat (versus those who did not). The small insignificant coefficient on the uninteracted Sit-D term shows that trained officers did not fire more in general (i.e., on those who did not pose a direct threat). These results suggest that trained officers shifted how they used their weapon, shooting more often in situations where it was appropriate for them to do so. This

suggests that the training did not make officers more passive in the face of direct threats, but rather enabled them to better respond with direct action when this was required.

Moreover, the results from the endline assessment are robust to alternative specifications. Table B11 shows that these results hold when we select covariates using a Double LASSO procedure. Table B12 additionally shows that the results also hold without the addition of any covariates.

Finally, note that some of the measures in the endline assessment might at first seem subject to experimenter demand. This could potentially be more of a concern for some of the measures discussed in Appendix B.1, such as officers' self-reported strategies for regulating stress or emotions. For instance, when Sit-D officers report using more breathing strategies to manage stress, experimenter demand could conceivably play a role in these responses. However, experimenter demand is unlikely to affect the key measures discussed above (such as the details that officers recall from the scenes, the specific explanations they generate for a subject's behavior, or how long they take to decide on an interpretation), since there is no obvious answer that trained officers "should" give. In addition, many of these items reflect what officers do, rather than simply what they say. The administrative outcomes from the field, described below, further capture officer behavior, and also cannot be subject to experimenter demand. Overall, the notable correspondence in effects across a range of measures suggests that our results cannot be attributed to experimenter demand alone.

Taken together, the results from the endline assessments highlight how Sit-D trains officers to think differently. Most importantly, the training improves how officers think through alternative interpretations. We now turn to data from the field to assess whether the training also affects behavior that leads to adverse policing outcomes.

## 4.2 Outcomes in the Field

To gauge impacts on field outcomes we present estimates of equation (2), in Table 3. All tables using administrative data follow a common structure. Each row represents a different

regression, corresponding to a different outcome. The first column shows the control mean for four months after the training (the key evaluation period when endline assessments were also conducted). The second column presents the treatment effect, which corresponds to estimates of $\beta$ from equation (2). The third and fourth columns present the standard errors and observed p-values respectively, while the fifth column presents the FDR-adjusted q-values.

We begin by discussing our two key adverse policing outcomes: uses of non-lethal force and our measure of discretionary arrests. The top row of Table 3 shows that Sit-D leads to significant and substantial reductions in the uses of force outcome. In the control group, there are 38 uses of non-lethal force for every 1,000 officers each month. The coefficient of -8.9 implies a 23% reduction in this outcome. As shown in Table B13, this effect stems from reductions in both the lowest level (Level 1) incidents as well as the higher level (Level 2) incidents. The Level 2 effect is more precisely estimated and implies a 30% reduction, while the Level 1 effect is qualitatively smaller, implying a 19% reduction. These results show that the reduction in uses of non-lethal force does not reflect Sit-D's impacts on the lowest levels of force alone.[21]

In Table B13 we also examine two other outcomes related to force incidents: subject injuries and tactics. We observe some evidence of a fall in officer-reported subject injury, though this effect should be taken as suggestive given measurement challenges in this variable. We also do not observe significant reductions in the use of force tactics versus other types of tactics, but this outcome is only defined conditional on a force incident occurring. See Appendix A.2 for a discussion of measurement issues and Appendix B.2 for a more in-depth discussion of these results.

The second row of Table 3 examines our second key adverse policing outcome, the num-

[21] Table B13 also examines other use of force aggregations including lethal force (Level 3) incidents. There are only 8 such incidents in our sample so we are not powered to examine this outcome individually. When Level 3 is combined with both Level 1 and 2 force incidents, the coefficient implies an 19% reduction in this outcome, with a p-value of .108. When Level 3 is combined with Level 2 incidents only, the low control mean indicates there are many fewer such incidents compared to Level 1 and 2 uses of force; thus the estimate is less precise but still implies a 20% reduction in this outcome.

ber of discretionary arrests. The results show that the training also leads to substantial reductions in this category of arrests (which, as discussed in the data section, likely arise when officers respond emotionally, out of irritation or frustration toward subjects). In the control group, there are 37 such discretionary arrests for every 1,000 officers each month. The coefficient of -8.5 implies a 23% reduction in this outcome.

Though the control means and effect sizes for uses of force and these types of discretionary arrests are similar in magnitude, it is important to note that these two outcomes do not necessarily stem from the same underlying incidents. For example, in the control group, the simple correlation between these two outcomes is .16 at the officer-month level. Moreover, 80% of uses of force take place in officer-months without any discretionary arrests, while 76% of discretionary arrests take place in officer-months with no uses of force. This underscores the fact that discretionary arrests can be made without employing force. And, force may be used for incidents unrelated to the charges underlying discretionary arrests. In short, these are two different outcomes, and the training leads to reductions in both.

The fall in uses of force and discretionary arrests observed among Sit-D trained officers may raise concerns that the training makes officers too passive, in ways that increases their risk of getting hurt on the job. To address this possibility, we also look at officer injuries. In particular, we analyze days of officer absence owing to injury on duty (IOD). We did not pre-specify examining this outcome in our PAP, but analyze it here given the importance of addressing potential concerns that the training may pose safety risks to officers. Note that IOD absences stem from a wide range of officer activities, not just use of force incidents. For example, if an officer falls or crashes their vehicle while rushing to a crime scene, or hurts their shoulder while forcing their way into a vacant apartment, these incidents will be logged as injuries (but are not use of force incidents).

The results in the third panel of Table 3 show that the training leads to a significant and substantial reduction in officer injuries. The control mean indicates that there are on average 1.2 IODs per officer per month. The coefficient of -.57 suggests that IODs are almost

half as large in the trained group as compared to the control group.

Another concern might be that the training produces less active officers who are engaged in fewer overall activities. If this were the case, uses of force could fall as a result of officers finding themselves in fewer situations that would potentially require force. To address this account, in the bottom panel of Table 3, we examine an index of overall officer activity, which comprises more than twelve different types of activities ranging from parking citations and curfew violations, to traffic stops, recovered guns and all other types of arrests not included in our subset of discretionary arrests (see "Other outcomes" under Section 3.2.2 for a complete list.) This result shows that Sit-D does not lead to any appreciable decreases in overall officer activity. In fact the coefficient is positive, suggesting an increase in activity, albeit imprecisely estimated.[22] It also makes clear that the reductions in officer injury cannot reflect lower levels of officer engagement.

Consistent with our predictions, we observe reductions in uses of force and discretionary arrests. However, we do not see reductions in overall officer activity.

To comprehensively analyze our administrative data, in Table B14, we examine two additional outcomes that are downstream responses to officers' actions: complaints (filed both by civilians and internally in CPD) and commendations and awards (see Appendix A.2 for further details on these variables). We find no significant effects on either outcome. It is possible that we do not detect effects here because these are noisy measures of officer performance relative to the direct actions taken by officers (such as arrests made or force deployed). For example, there may be political or bureaucratic factors that guide why some individuals are awarded commendations or shielded from department-initiated complaints. It is also possible that Sit-D does not change officer behavior in ways that induce further downstream responses from others (such as whether they file complaints against officers).

[22]While the p-value is insignificant at conventional levels, the q-value is significant at the 10% level. As discussed in the code implementing FDR adjustment for Anderson (2008), q-values can be lower than p-values. For example, this can occur when multiple hypotheses are rejected (or when the un-adjusted p-values are relatively low), because if there are multiple true rejections, several false rejections can also be tolerated.

**Robustness Checks on Field Outcomes.** In this section we check the robustness of our main results to alternate specifications and variable definitions.

Our discretionary arrests category constitutes a small subset of arrests (covering 1.3% of all arrests) that we thought were most likely to fall in response to the training. To gauge if Sit-D also reduces other types of arrests over which officers might have discretion, we use a different measure of discretionary arrests based on Rivera and Ba (2022), which consists of arrests for "non-index" crimes under the FBI's classification. We did not specify analyzing this outcome in our PAP, but examine it for robustness. Non-index arrests are typically perceived to be for low-level charges of what are called "victimless crimes." However, they are fairly broad in scope, accounting for 74% of all arrests in our sample, and some charges seem more serious than others.

Table B15 shows that Sit-D leads to a 9% reduction in the sum of all non-index crime arrests, though the effect is not significant at conventional levels, with a p-value of .12. However, the table also shows that this noisy effect reflects considerable heterogeneity across various FBI charge categories. Sit-D in fact leads to a significant *increase* in arrests for one relatively serious non-index category: criminal sexual abuse. In contrast, it leads to significant reductions in arrests for both gambling and municipal code violations. Most of these "gambling" charges are for individuals playing games, like shooting dice on the street (97% of the gambling arrests in our sample are for the category "playing game of chance"). In addition, municipal code violations are for crimes like littering and riding a bicycle on the street, which are similar in spirit to our discretionary arrests measures in that making these arrests may have small effects on public safety. The reduction in these low-level non-index crime arrests appears to be consistent with why we observe a reduction in discretionary arrests—these non-index arrests might also arise in ambiguous situations where greater deliberation could lead officers to adopt a less severe response.

We next check the robustness of our results to a different focal period. We conducted our endline assessments four months after the training, and we evaluate field outcomes over

this focal period in the main section of the paper. But, in Table B16, we ask what the observed effects would have been if the focal period were instead instead three months after the training. We find similar sized and, if anything, qualitatively larger reductions in both of the adverse policing outcomes: We find a 25% reduction in uses of force and a 29% reduction in discretionary arrests over this alternate period. In addition, we continue to see substantial reductions in officer injuries while total officer activities remain unchanged.

As with the endline assessment results, our field results are also insensitive to specific controls. Our baseline specifications incorporate, in all outcome regressions, a set of common covariates including officer characteristics and baseline administrative data. Panel A of Table B17 shows that the results remain unchanged if we instead incorporate covariate sets into each outcome regression using the LASSO double-selection procedure. Panel B of this table also verifies that results hold without the addition of any covariates. As expected, the estimates are more precise with controls, but our results are not dependent on any one approach. We also verify that our results are not sensitive to randomly allocating control officers to post-training periods (see Appendix B.2 and Table B18 for a specification in which we use multiple post-training periods for each control officer).

In the appendix, we also estimate a difference-in-differences (DID) model. This approach will have less power than our main specification since there is low autocorrelation in key outcomes (McKenzie, 2012).[23] Consistent with this, Figure B1 shows that the DID estimates are indeed less precise than the baseline estimates (from Table 3). However, as discussed in Appendix B.2, we cannot reject the null hypothesis that the estimates from these models are the same based on tests from Seemingly Unrelated Estimation.

**Heterogeneity by Officer and District Characteristics.** In the next two subsections, we discuss heterogeneous effects of the training. First, we examine differences based on

[23]For example, autocorrelation in uses of non-lethal force and discretionary arrests are .335 and .295 respectively over the baseline and post-training period. This likely reflects major changes that affected the policing environment, including the COVID-19 pandemic and the killing of George Floyd, which sparked nation-wide protests against policing practices.

characteristics of the officers and the districts in which they are employed.

Since these characteristics may be correlated with other factors that shape adverse policing outcomes, we do not advance a causal interpretation of these analyses, but rather use them to provide suggestive evidence on which types of officers may benefit most from the training.

We see clear patterns of heterogeneity based on officer experience: Table B19 shows that Sit-D leads to larger reductions in both adverse policing outcomes among officers who have been on the job for fewer years. In terms of officer demographics, we do not see significant differential effects based on the race of the officer, but do observe significantly larger responses to the training among male officers. Since inexperienced officers and male officers have higher uses of force at baseline (Ba et al., 2021), these findings suggest that Sit-D has greater impact among officers who face worse starting points, for whom training needs may be greater. Table B21, which examines heterogeneity by baseline measures of the adverse policing outcomes corroborates this interpretation. Appendix B.2 provides a more detailed discussion of these results.

Finally, we consider if the benefits of the training are localized to places where officers face relatively little risk (in Table B22). However, we do not observe significant differential effects based on crime rates in the officer's district of employment. This suggests that the benefits of Sit-D are widespread across different types of risk environments officers might face.

**Heterogeneity by Subject's Race.** Next, we consider heterogeneity based on the race of the subjects. Table 4 presents discretionary arrests separately for Black subjects and other subjects. The top two rows of the table show that the reduction in this outcome is driven by arrests of Black subjects specifically. The remaining rows of this table show that this same pattern holds for all arrests, as well as other arrests that were not pre-specified as discretionary. Table B23 also verifies that the null effect on subjects of other races holds

when this category is disaggregated into Hispanic, White, and other races.[24]

The implied effects from Table 4 are substantial: Sit-D leads to a 11 % fall in the arrests of Black civilians, but a 3% fall in the arrests of other subjects. The difference is especially large for discretionary arrests: the training leads to a 28% reduction for Black civilians while directionally (but insignificantly) increasing these arrests by 7% for other subjects.

Tests of equality from Seemingly Unrelated Regressions (SURs) verify that the effects on the number of arrests for Black versus other subjects are significantly different from one another, for all three arrest categories.[25]

These results establish a disparate impact of the training on Black civilians, who are arrested at much higher rates than other subjects—6 times higher for discretionary arrests, and 3 times higher for other types of arrests (see column 1 of Table 4). Thus, Sit-D could lead to a larger reduction in the number of Black arrests and produce a disparate impact even if the training uniformly changes how officers engage with civilians of all races.

Does the training also differentially change how officers engage with Black subjects? In other words, is there a differential treatment effect for this group, above and beyond disparate impact? To consider this, we examine the number of arrests relative to the race-specific control mean. Table B24 repeats the analysis from Table 4, with Z-scores of the arrest variables (in which we subtract the control mean and scale by the control standard deviation). These results show that Sit-D leads to a larger reduction in the arrests of Black subjects even relative to the higher rate at which they are arrested in the control group. While this pattern holds qualitatively for all three arrest categories, SUR tests show that the race effect is significantly larger for discretionary arrests, specifically.[26] Thus the training appears to produce differential treatment effects for Black civilians when officers have discretion in how

[24]The differential impact we observe on Black subjects is consistent with Rivera (2022), which finds reductions in low-level arrests for Black subjects specifically, when examining race-based peer effects among police officers.

[25]We reject the null hypothesis that the effects are the same for Black subjects and other subjects with a p-value of .040 for all arrests, .030 for discretionary arrests and .047 for other arrests.

[26]The p-values from these SUR tests of equality across black and other subjects are .17 for all arrests, .09 for discretionary arrests, and .19 for other arrests.

to resolve the situation. These findings are notable because the Sit-D curriculum did not contain modules focusing on either implicit or explicit racial biases in policing. But, to the extent that officers' initial impressions are more biased, or more prone to assumptions in situations involving Black civilians, then by getting officers to consider multiple interpretations of a situation, Sit-D might reduce the effect of these first impressions on officers' behavior (Axt and Lai, 2019).

Additional analyses help unpack Sit-D's effects on racial disparities in arrests, and their implications. These are discussed in detail in Appendix B.2, but a few points are worth highlighting. In Table B25, we observe similarly sized reductions in racial disparities for multiple levels of arrests (e.g., arrests for non-index, property, and violent crimes). Moreover, as shown in Table B26 the reduction in "other arrests" of Black subjects corresponds to an overall reduction in the "other arrests" category. This reflects the fact that 77% of all such arrests in the control group are arrests of Black subjects.

However, this decline does not correspond to officers becoming less active overall, as shown by the index of activities. Moreover, given the differing objectives that police departments might have, it is difficult to say whether this drop in "other arrests" is desirable or undesirable. As other scholars have noted, a reduction in arrests of Black subjects should not be taken as a reduction in productivity from a public safety standpoint (Rivera and Ba, 2022; Ba et al., 2022; Lum and Nagin, 2017). Indeed, in our data, suggestive analyses indicate that these reductions in arrests do not correspond to increases in crime (see Table B27). Thus, it appears that Sit-D alters racial patterns of arrests without necessarily increasing crime or reducing overall police activity.

**Examining Spillovers.** Next, we examine potential spillovers, which will lead us to underestimate the treatment effect. We discuss this briefly here but provide a fuller description in Appendix B.2. Drawing on the approach of Ba et al. (2021), we utilize variation in the ratio of Sit-D officers to *total* officers assigned to unit-watches at the time of randomization.

We then regress our main field outcomes on Sit-D and its interaction with this ratio.

The results in Table B28 indicates there are substantial spillovers on the two key adverse policing outcomes. The coefficient on Sit-D x Ratio is negative, indicating that a larger fraction of Sit-D officers reinforces the training's effect in reducing both uses of force and discretionary arrests. The magnitudes imply that increasing the Sit-D ratio from its mean (17%) to its max (35%) elevates the reduction in uses of force from 27 percent to 79 percent; and elevates the reduction in discretionary arrests from 23 percent to 78 percent (relative to the overall control mean). Since a larger share of treated officers reinforce the training's effect, these patterns suggest that the presence of untrained officers lowers Sit-D's effect among trained officers.

**Effects in Later Periods.** Since skills acquired through training may perish over time, we next examine the period over which Sit-D's effects are sustained. To do so, we pool together 12 months of administrative data and examine effects over 5-8 and 9-12 months after the training. We estimate equation (3) and plot the coefficients and 90% Confidence Intervals (along with p-values and q-values) in Figure 2.

These results suggest that the effects diminish over the course of the year, though given the size of the confidence intervals, it is not clear exactly when this happens for all outcomes. For uses of non-lethal force, the estimates for both of the additional periods are individually insignificant. However, they are also not significantly different from the effect in the focal period one to four months after the training. Specifically, we fail to reject the null hypothesis that the coefficient estimates for the three periods are equal to one another (with p-value of .37).

For discretionary arrests, the estimate in months 5-8 is significant and implies a 26% reduction over this period, though the estimate is sensitive to FDR adjustment; while the effect in months 9-12 is insignificant. However, again, the confidence intervals are large and we cannot reject the null hypothesis that the coefficient estimates for all three periods are

equal to one another (with a p-value of .22).

In contrast, Figure 2 shows that the effects on officer injuries do differ significantly across the three periods, and that the positive effect on officer activities index is significantly smaller in the latter two periods relative to the first period (with p-values of .09 and .10). Thus, while the timing of fade-out for the two adverse policing outcomes is unclear, the estimates on these other outcomes more clearly indicate that fade-out begins to occur five months after the training. As such, it might be beneficial to introduce refresher trainings during this time interval to reinforce Sit-D's effects.

There are various possible reasons why this fade-out might occur. First, in light of the negative spillover effects discussed above, it is possible that increased interactions with untreated officers contribute to lowering the effectiveness of the training over time. If this is the case, we should observe a higher fraction of Sit-D officers sustaining the effects of the training in later periods. Table B29 presents *some* evidence consistent with this account. For both months 5-8 and 9-12 the coefficient on Sit-D x Ratio is negative for both adverse policing outcomes. However this effect is more precisely estimated for the first of these two periods, suggesting that spillovers alone may not account fully for the fade-out. This indicates that the treatment effectiveness itself may also diminish over time. This type of diminishing effect is common among debiasing and cognitive training interventions (Korteling et al., 2021), for a number of reasons. Some research suggests that structural neural characteristics are responsible for cognitive biases in how people search for and use information when making decisions (Korteling et al., 2018), which would make it harder to debias people in the long run without refresher trainings. More generally, habits (including, perhaps, habits of thought) are often difficult to change because our mental systems evolved in environments that did not require relatively rapid changes (Poldrack, 2021).

Since treatment effects are strongest in the first four-month interval, and appear to diminish thereafter, Table B30 additionally examines the pooled effect twelve months after the training. We observe a 12% reduction in uses of force and a 18% reduction in discretionary

arrests over this aggregate period. The effect is significant at the 5% and 10% level in p-value and q-value, respectively, for discretionary arrests; and at the 10% level in q-value for uses of non-lethal force. Thus, though the effects weaken over time, there is still *some* sustained effect on adverse policing outcomes over the aggregate 12-month period after the training. Overall, however, the pattern of results suggest that it will be necessary to re-train officers during the year to strongly sustain the training's effect over this duration.

# 5 Discussion

Policing takes place in cognitively demanding situations. We suggest that in the face of these cognitive demands, officers might act without fully considering alternative interpretations of the situations they encounter. And this can contribute to adverse outcomes.

Importantly, we find that it is possible to mitigate these adverse outcomes by training officers to manage the cognitive demands of policing. Our endline assessments show that officers trained in Sit-D were better able to think through ambiguous situations. And, in the field, trained officers used less force and made fewer discretionary arrests, while also experiencing fewer injuries. These results show that officer behavior is remarkably malleable and responsive to this type of training.

Moreover, our findings also show that the training helps reduce racial disparities in policing. Specifically, trained officers arrested fewer Black civilians overall. This is notable given how persistent these disparities have been (Correll et al., 2007; Fryer, 2019; Rozema and Schanzenbach, 2019; Goncalves and Mello, 2021; Hoekstra and Sloan, 2022). While departments often turn to implicit bias training, there is little evidence that such trainings are effective (Worden et al., 2020; Lai and Lisnek, 2023). As discussed above, perhaps a more effective route to reducing racial disparities in policing is to use cognitive training that makes officers more deliberative.

The cost of Sit-D per officer trained ($807-$864) appears to be roughly on the order

of other existing police trainings that use similar equipment (See Appendix C for further details on these cost calculations). Yet while we have evidence of Sit-D's effectiveness, we do not have corresponding evidence around the impact of those other trainings. The benefits of Sit-D are unfortunately harder to measure since many are likely to be "non-market" benefits, such as reductions in the physical and psychological costs stemming from fewer uses of force and low-value arrests. Reductions in adverse policing outcomes also affect broad societal responses, such as social unrest, along with trust in (and cooperation with) law enforcement. However, even the benefits from reduced officer injuries alone exceed the costs of training. As discussed in Appendix C, we find that Sit-D would save $1057 in personnel costs per officer trained in the four months post-training. Given the much larger range of potential benefits (at a cost comparable to other trainings), Sit-D appears to be a promising lever for police departments to draw on.

As such, developing a further understanding of how cognitive demands affect policing (and how to train for these demands) is fertile ground for future work. For instance, it will be important to examine how peer effects interact with such training. Prior research suggests that officers exert considerable influence over each other's behavior (Getty et al., 2016; Adger et al., 2022), and these types of peer effects shape key officer outcomes (Holz et al., 2023; Rivera, 2022). Indeed, we present suggestive evidence showing that spillovers may lead us to understate the true effect of the training. At the same time, the presence of these spillovers indicate that training a larger fraction of officers could help reinforce the principles of the training, leading to more sustained effects.

There also remain pragmatic questions about how to deliver such a training most effectively. For instance, what is the ideal intensity of the training both in terms of total hours as well as how those hours are distributed over weeks or months? Or, how often are refresher trainings needed to maintain these effects? Answering these questions will require iterations on training configuration alongside more precise estimates of the durability of effects in the field, which will help translate the principles outlined here into scalable trainings.

Ultimately, the concepts and training presented here offer an important complement to existing perspectives on how to reduce adverse policing outcomes. The view that problem officers are responsible for these outcomes has spurred recent research on the benefits of early warning systems to detect these officers (Chalfin and Kaplan, 2021; Sierra-Arévalo and Papachristos, 2021). Meanwhile, the view that bad regulations are to blame has given rise to work on how department policy (Mummolo, 2018) and accountability (Prendergast, 2021; Rivera and Ba, 2022) affects officer behavior. Our perspective suggests that—given the demands inherent in policing—there may also be benefits to teaching officers how to think critically during stressful situations, without necessarily telling them exactly how to respond. Officers might be better able to adapt and respond to a variety of situations if they are trained to meet the cognitive demands of policing.

# References


Adger, C., M. Ross, and C. Sloan (2022). The Effect of Field Training Officers on Police Use of Force. *Working Paper*.

Alexander, K. L., S. Rich, and H. Thacker (2022). The Hidden Billion Dollar Cost of Repeated Police misconduct. The Washington Post.

Anderson, M. L. (2008). Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects. *Journal of the American Statistical Association 103*(484), 1481–1495.

Ang, D. (2020, 09). The Effects of Police Violence on Inner-City Students. *The Quarterly Journal of Economics 136*(1), 115–168.

Axt, J. and C. K. Lai (2019). Reducing discrimination: a bias versus noise perspective. *Journal of Personality and Social Psychology 117*, 26–49.

Ba, B., P. Bayer, N. Rim, R. Rivera, and M. Sidibe (2022, May). Police Officer Assignment and Neighborhood Crime. Working Paper 29243, National Bureau of Economic Research.

Ba, B. A. (2020). Going the Extra Mile: The Cost of Complaint Filing, Accountability, and Law Enforcement Outcomes in Chicago.

Ba, B. A., D. Knox, J. Mummolo, and R. Rivera (2021). The Role of Officer Race and Gender in Police-Civilian Interactions in Chicago. *Science 371*(6530), 696–702.

Banerjee, A., R. Chattopadhyay, E. Duflo, D. Keniston, and N. Singh (2021, 02). Improving Police Performance in Rajasthan, India: Experimental Evidence on Incentives, Managerial Autonomy, and Training. *American Economic Journal: Economic Policy 13*(1), 36–66.

Belloni, A., V. Chernozhukov, and C. Hansen (2013, 11). Inference on Treatment Effects after Selection among High-Dimensional Controls†. *The Review of Economic Studies 81*(2), 608–650.

Benjamini, Y., A. M. Krieger, and D. Yekutieli (2006, 09). Adaptive Linear Step-up Procedures that control the False Discovery Rate. *Biometrika 93*(3), 491–507.

Bhatt, M. P., S. B. Heller, M. Kapustin, M. Bertrand, and C. Blattman (2023, January). Predicting and Preventing Gun Violence: An Experimental Evaluation of READI Chicago. Working Paper 30852, National Bureau of Economic Research.

Blattman, C., J. C. Jamison, and M. Sheridan (2017, 04). Reducing Crime and Violence: Experimental Evidence from Cognitive Behavioral Therapy in Liberia. *American Economic Review 107*(4), 1165–1206.

Bodenhausen, G., L. Sheppard, and G. Kramer (1994, 01). Negative Affect and Social Judgment: The Differential Impact of Anger and Sadness. *European Journal of Social Psychology 24*, 45 – 62.

Bor, J., A. S. Venkataramani, D. R. Williams, and A. C. Tsai (2018, 06). Police Killings and Their Spillover Effects on the Mental Health of Black Americans: A Population-based, Quasi-experimental Study. *Lancet 392*(10144), 302–310.

Bordalo, P., K. Coffman, N. Gennaioli, and A. Shleifer (2016, 07). Stereotypes. *The Quarterly Journal of Economics 131*(4), 1753–1794.

Bordalo, P., N. Gennaioli, and A. Shleifer (2013, May). Salience and Asset Prices. *American Economic Review 103*(3), 623–28.

Bordalo, P., N. Gennaioli, and A. Shleifer (2015). Salience Theory of Judicial Decisions. *The Journal of Legal Studies 44*(S1), S7–S33.

Brenan, M. (2023). Americans More Critical of U.S. Criminal Justice System. https://news.gallup.com/poll/544439/americans-critical-criminal-justice-system.aspx.

Brownback, A., A. Imas, and M. A. Kuhn (2023, 02). Behavioral Food Subsidies. *The Review of Economics and Statistics*, 1–47.

Canales, R., M. Magaña, J. F. Santini, and A. C. Maus (2020). Assessing the Effectiveness of Procedural Justice Training for Police Officers: Evidence from the Mexico City Police. *Working Paper*.

Chaiken, S. (1980). Heuristic Versus Systematic Information Processing and the Use of Source Versus Message Cues in Persuasion. *Journal of Personality and Social Psychology 39*(5), 752.

Chalfin, A., B. Hansen, E. K. Weisburst, and J. Williams, Morgan C. (2022, June). Police Force Size and Civilian Race. *American Economic Review: Insights 4*(2), 139–58.

Chalfin, A. and J. Kaplan (2021). How Many Complaints Against Police Officers Can Be Abated by Incapacitating A Few 'Bad Apples?'. *Criminology & Public Policy 20*(2), 351–370.

Chalfin, A. and J. McCrary (2018, 03). Are U.S. Cities Underpoliced? Theory and Evidence. *The Review of Economics and Statistics 100*(1), 167–186.

Chen, T. H. Y., P. McLachlan, and C. J. Fariss (2021, 10). Exposure to Discretionary Arrests Increases Support for Anti-Police Protests. *Working Paper*.

Chicago Police Department (2021). CPD General Order. `http://directives.chicagopolice.org/#directive/public/6610`. Accessed: 2022-06-13.

Cho, S., F. Gonçalves, and E. Weisburst (2022). Do Police Make Too Many Arrests? The Effect of Enforcement Pullbacks on Crime. *IZA Discussion Paper No. 14907*.

Correll, J., B. Park, C. M. Judd, B. Wittenbrink, M. S. Sadler, and T. Keesee (2007, 06). Across the Thin Blue Line: Police Officers and Racial Bias in the Decision to Shoot. *Journal of Personality and Social Psychology 92*(6), 1006–1023.

Desmond, M., A. V. Papachristos, and D. S. Kirk (2016). Police Violence and Citizen Crime Reporting in the Black Community. *American Sociological Review 81*(5), 857–876.

Dunning, D., D. Griffin, J. Milojkovic, and L. Ross (1990, 05). The Overconfidence Effect in Social Prediction. *Journal of Personality and Social Psychology 58*, 568–81.

Engel, R. S., N. Corsaro, G. T. Isaza, and H. D. McManus (2022). Assessing the Impact of De-escalation Training on Police Behavior: Reducing Police Use of Force in the Louisville, KY Metro Police Department. *Criminology & Public Policy 21*(2), 199–233.

Engel, R. S., H. D. McManus, and T. D. Herold (2020). Does De-escalation Training Work? *Criminology & Public Policy 19*(3), 721–759.

Enke, B. (2020, 04). What You See Is All There Is. *The Quarterly Journal of Economics 135*(3), 1363–1398.

Fagan, J. A. and A. D. Campbell (2020). Race and Reasonableness in Police Killings. Available at: Columbia Law School Scholarship Archive.

Fearon, J. D. (2019). Coups, Police Shootings, and Nuclear War. *Paper presented at the 2019 Annual Meetings of the American Political Science Association, Washington DC August 29-September 1*.

Fischhoff, B., P. Slovic, and S. Lichtenstein (1978, 05). Fault Trees: Sensitivity of Estimated Failure Probabilities to Problem Representation. *Journal of Experimental Psychology: Human Perception and Performance 4*, 330–344.

Fiske, S. T. and S. L. Neuberg (1990). A Continuum of Impression Formation, from Category-Based to Individuating Processes: Influences of Information and Motivation on Attention and Interpretation. In *Advances in Experimental Social Psychology*, Volume 23, pp. 1–74. Elsevier.

Fryer, R. G. (2019). An Empirical Analysis of Racial Differences in Police Use of Force. *Journal of Political Economy 127*(3), 1210–1261.

Gabaix, X. (2019). Chapter 4 - Behavioral inattention. In B. D. Bernheim, S. DellaVigna, and D. Laibson (Eds.), *Handbook of Behavioral Economics - Foundations and Applications 2*, Volume 2 of *Handbook of Behavioral Economics: Applications and Foundations 1*, pp. 261–343. North-Holland.

Getty, R. M., J. L. Worrall, and R. G. Morris (2016). How Far From the Tree Does the Apple Fall? Field Training Officers, Their Trainees, and Allegations of Misconduct. *Crime & Delinquency 62*(6), 821–839.

Gilbert, D. T., B. W. Pelham, and D. S. Krull (1988). On Cognitive Busyness: When Person Perceivers Meet Persons Perceived. *Journal of Personality and Social Psychology 54*(5), 733.

Goncalves, F. and S. Mello (2021, 05). A Few Bad Apples? Racial Bias in Policing. *American Economic Review 111*(5), 1406–41.

Grossi, D. (2017, 08). Police firearms training: How often should you be shooting? Police1 [Online; posted 23-June-2011; updated 11-August-2017].

Harcourt, B. E. and J. Ludwig (2006). Broken Windows: New Evidence from New York City and a Five-City Social Experiment. *The University of Chicago Law Review 73*(1), 271–320.

Haseman, J., K. Zaiets, M. Thorson, C. Procell, G. Petras, and S. J. Sullivan (2020, 06). Tracking Protests across the USA in the Wake of George Floyd's Death. USA TODAY [Online; posted 3-June-2020].

Heller, S. B., A. K. Shah, J. Guryan, J. Ludwig, S. Mullainathan, and H. A. Pollack (2017, 10). Thinking, Fast and Slow? Some Field Experiments to Reduce Crime and Dropout in Chicago*. *The Quarterly Journal of Economics 132*(1), 1–54.

Hoekstra, M. and C. Sloan (2022, 03). Does Race Matter for Police Use of Force? Evidence from 911 Calls. *American Economic Review 112*(3), 827–60.

Holz, J. E., R. G. Rivera, and B. A. Ba (2023). Peer Effects in Police Use of Force. *American Economic Journal: Economic Policy 15*(2), 256–291.

Imas, A., M. A. Kuhn, and V. Mironova (2022). Waiting to choose: The role of deliberation in intertemporal choice. *American Economic Journal: Microeconomics 14*(3), 414–40.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Number 6. Harvard University Press.

Jones, J. M. (2022, 07). Confidence in U.S. Institutions Down; Average at New Low. Gallup.com [Online; posted 5-May-2022].

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.

Kassam, K. S., K. Koslov, and W. B. Mendes (2009). Decisions Under Distress: Stress Profiles Influence Anchoring and Adjustment. *Psychological Science 20*(11), 1394–1399. PMID: 19843261.

Kling, J. R., J. B. Liebman, and L. F. Katz (2007). Experimental Analysis of Neighborhood Effects. *Econometrica 75*(1), 83–119.

Korteling, J. E., A.-M. Brouwer, and A. Toet (2018). A Neural Network Framework for Cognitive Bias. *Frontiers in Psychology 9*.

Korteling, J. H., J. Y. J. Gerritsma, and A. Toet (2021). Retention and Transfer of Cognitive Bias Mitigation Interventions: A Systematic Literature Study. *Frontiers in Psychology 12*.

Lai, C. K. and J. A. Lisnek (2023). The impact of implicit bias-oriented diversity training on police officers' beliefs, motivations, and actions. In press.

Lerner, J. S., Y. Li, P. Valdesolo, and K. S. Kassam (2015). Emotion and Decision Making. *Annual Review of Psychology 66*(1), 799–823. PMID: 25251484.

Levitt, S. D. (1998, December). Juvenile Crime and Punishment. *Journal of Political Economy 106*(6), 1156–1185.

Lum, C. and D. S. Nagin (2017). Reinventing American Policing. *Crime and Justice 46*, 339–393.

McKenzie, D. (2012). Beyond baseline and follow-up: The case for more T in experiments. *Journal of Development Economics 99*(2), 210–221.

McLean, K., S. E. Wolfe, J. Rojek, G. P. Alpert, and M. R. Smith (2020). Randomized controlled trial of social interaction police training. *Criminology & Public Policy 19*(3), 805–832.

Moreno-Medina, J., A. Ouss, P. Bayer, and B. Ba (2024). Officer-Involved: The Media Language of Police Killings. *Quarterly Journal of Economics*. Conditionally accepted.

Mummolo, J. (2018). Modern Police Tactics, Police-Citizen Interactions, and the Prospects for Reform. *The Journal of Politics 80*(1), 1–15.

Nickerson, R. S. (1998). Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology 2*(2), 175–220.

Owens, E., D. Weisburd, K. L. Amendola, and G. P. Alpert (2018). Can You Build a Better Cop? *Criminology & Public Policy 17*(1), 41–87.

Payne, J. W., J. R. Bettman, and E. J. Johnson (1993). *The Adaptive Decision Maker*. Cambridge University Press.

Petty, R. E. and J. T. Cacioppo (1986). The Elaboration Likelihood Model of Persuasion. In *Communication and Persuasion*, pp. 1–24. Springer.

Poldrack, R. (2021). *Hard to Break: Why Our Brains Make Habits Stick*. Princeton University Press.

Prendergast, C. (2021). 'Drive and Wave': The Response to LAPD Police Reforms After Rampart. *University of Chicago, Becker Friedman Institute for Economics Working Paper No. 2021-25*, 371–381.

Rad, A. N., D. S. Kirk, and W. P. Jones (2023). Police Unionism, Accountability, and Misconduct. *Annual Review of Criminology 6*(1).

Rivera, R. (2022). The Effect of Minority Peers on Future Arrest Quantity and Quality. *Available at SSRN 4067011*.

Rivera, R. and B. A. Ba (2022, 02). The Effect of Police Oversight on Crime and Allegations of Misconduct: Evidence from Chicago. *Working Paper*.

Rosenbaum, D. P. and D. S. Lawrence (2017, 09). Teaching procedural justice and communication skills during police–community encounters: Results of a randomized control trial with police recruits. *Journal of Experimental Criminology 13*(3), 293–319.

Rozema, K. and M. Schanzenbach (2019, 05). Good Cop, Bad Cop: Using Civilian Allegations to Predict Police Misconduct. *American Economic Journal: Economic Policy 11*(2), 225–68.

Schaefer, B. P. and T. Hughes (2019, 08). Examining Judicial Pretrial Release Decisions: The Influence of Risk Assessments and Race. *Criminology, Criminal Justice, Law & Society 20*(2).

Schuck, A. M. and D. P. Rosenbaum (2005, 12). Global and Neighborhood Attitudes Toward the Police: Differentiation by Race, Ethnicity and Type of Contact. *Journal of Quantitative Criminology 21*(4), 391–418.

Schwartz, J. C. (2016). How Governments Pay: Lawsuits, Budgets, and Police Reform. *UCLA Law Review 63*(5), 1144.

Shah, A. and D. M. Oppenheimer (2008). Heuristics Made Easy: An Effort-Reduction Framework. *Psychological Bulletin 134*(2).

Shaklee, H. and B. Fischhoff (1982, 11). Strategies of Information Search and Causal Analysis. *Memory & Cognition 10*(6), 520–530.

Sierra-Arévalo, M. and A. Papachristos (2021). Bad Apples and Incredible Certitude. *Criminology & Public Policy 20*(2), 371–381.

Simon, H. A. (1955, 02). A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics 69*(1), 99–118.

Skogan, W. G., M. Van Craen, and C. Hennessy (2015, 09). Training Police for Procedural Justice. *Journal of Experimental Criminology 11*(3), 319–334.

Walker, S., G. P. Alpert, and D. J. Kennedy (2001, 07). Early Warning Systems: Responding to the Problem Police Officer. *Working Paper*.

Weisburd, D., C. W. Telep, H. Vovak, T. Zastrow, A. A. Braga, and B. Turchan (2022). Reforming the Police through Procedural Justice Training: A Multicity Randomized Trial at Crime Hot Spots. *Proceedings of the National Academy of Sciences 119*(14), e2118780119.

Weitzer, R. and S. Tuch (2004, 08). Race and Perceptions of Police Misconduct. *Social Problems - SOC PROBL 51*.

Williamson, V., K.-S. Trump, and K. L. Einstein (2018). Black Lives Matter: Evidence that Police-Caused Deaths Predict Protest Activity. *Perspectives on Politics 16*(2), 400–415.

Wood, G., T. R. Tyler, and A. V. Papachristos (2020). Procedural Justice Training Reduces Police Use of Force and Complaints against Officers. *Proceedings of the National Academy of Sciences 117*(18), 9815–9821.

Wood, G., T. R. Tyler, and A. V. Papachristos (2021). Correction for Wood et al., Procedural justice training reduces police use of force and complaints against officers. *Proceedings of the National Academy of Sciences 118*(27), e2110138118.

Worden, R. E., S. J. McLean, R. S. Engel, H. Cochran, N. Corsaro, D. Reynolds, C. J. Najdowski, and G. T. Isaza (2020, 07). The Impacts of Implicit Bias Awareness Training in the NYPD. *Working Paper*.

# Tables and Figures

Table 1: Considering Alternative Interpretations

| | CM (1) | Sit-D (2) | SE (3) | p-value (4) | q-value (5) |
|---|---|---|---|---|---|
| **Panel A: Alternative Interpretations of a Subject's Actions** | | | | | |
| Total explanations | 3.215 | -0.013 | 0.077 | 0.862 | 0.606 |
| Explanations from multiple categories | 0.667 | 0.041 | 0.023 | 0.075* | 0.099* |
| At least one explanation - assistance category | 0.578 | 0.058 | 0.025 | 0.019** | 0.063* |
| At least one explanation - enforcement category | 0.624 | -0.008 | 0.024 | 0.742 | 0.590 |
| At least one explanation - other category | 0.676 | 0.000 | 0.024 | 0.989 | 0.687 |
| **Panel B: Processing Information and Forming Interpretations** | | | | | |
| Alternative Features Index (both tasks) | - | 0.101 | 0.032 | 0.001*** | 0.012** |
| Confirming Features Index (both tasks) | - | -0.014 | 0.032 | 0.675 | 0.572 |
| Criminal Interpretations Index (both tasks) | - | -0.052 | 0.025 | 0.040** | 0.083* |
| Decision Time Index (both tasks) | - | -0.062 | 0.032 | 0.052* | 0.083* |
| Processing Time Index (officer-timed task) | - | -0.021 | 0.044 | 0.627 | 0.572 |
| **Panel C: Use of Force in Dynamic Situations** | | | | | |
| Change – perceived threat & force assessment (index) | - | -0.077 | 0.039 | 0.048** | 0.083* |
| Appropriate actions (index) | - | 0.070 | 0.037 | 0.057* | 0.083* |
| Inappropriate actions (index) | - | -0.007 | 0.033 | 0.824 | 0.606 |

**Notes.** This table shows the effect of Sit-D training on the consideration of alternative interpretations (as measured by three tasks in the endline assessment), based on estimating equation (1). The top panel shows how officers described the subject in the Driver's Actions Task, the middle panel shows how offcers processed information and formed interpretations in the Pictures Task, and the bottom panel shows how offcers responded to use of force scenarios. Each row is a different regression. One observation is included for each offcer (N=1,582 for the top panel; N=1,669 for the middle and bottom panels). All regressions include stratum fixed effects and additional offcer-level covariates (race, gender, experience; as well as baseline values of discretionary arrests, offcer injuries and an index of offcer activities, key outcomes from the administrative data measured similarly at baseline and endline). Column (1) shows the control mean for each outcome (blank for mean effect indices). Column (2) shows the coeffcients on the Sit-D indicator. Column (3) shows robust standard errors. Column (4) shows the observed p-value. Column (5) shows the multiple-inference corrected q-values that adjust for the false discovery rate across outcomes in a family. All outcomes in this table are part of the Navigating Cognitively Demanding Situations Family. *** is significant at the 1% level, ** at the 5% level, and * at the 10% level.

Table 2: Performance in the FOS

**Panel A: Movement and Communication in the FOS**

| | Sit-D | | | |
|---|---|---|---|---|
| | Coef | SE | p-value | q-value |
| Did the officer communicate with the person? (index) | 0.127 | 0.029 | <0.001*** | 0.001*** |
| Did the officer give verbal direction/ commands to the person? (index) | 0.145 | 0.028 | <0.001*** | 0.001*** |
| Did the officer radio dispatch? (index) | 0.407 | 0.033 | <0.001*** | 0.001*** |
| Did the officer freeze during the scenario? (index) | -0.069 | 0.037 | 0.058* | 0.049** |
| Did the officer kneel or move to cover/ concealment? (index) | 0.040 | 0.033 | 0.230 | 0.131 |

**Panel B: Shooting in the FOS**

| | Sit-D | | | Direct Threat | | | Sit-D × Direct Threat | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Coef | SE | p-value | Coef | SE | p-value | Coef | SE | p-value | q-value |
| Shooting in the FOS | 0.007 | 0.019 | 0.713 | 0.601 | 0.016 | 0.000*** | 0.050 | 0.022 | 0.020** | 0.026** |

**Notes.** This table shows the effect of Sit-D training on officers' performance in the FOS exercises (measured in the endline assessment). The top panel shows the training's effects on movement and communication in the FOS, based on estimating equation (1). Each row is a different regression. One observation is included for each officer. N=1,611. From left to right, the columns show the coefficients on the Sit-D indicator, robust standard errors, the observed p-value, and multiple-inference corrected q-values that adjust for the false discovery rate across outcomes in a family. All outcomes in the top panel are part of the Officer Performance in the FOS Family. The bottom panel shows the training's effects on officers' decisions to shoot subjects in the FOS. One observation is included for each scenario completed by each officer. N=4,377. Direct Threat is an indicator for scenarios in which the subjects pose a direct threat. This panel shows the coefficients, robust standard errors, and observed p-values on each term. The last column shows the multiple-inference corrected q-value for the interaction term of Sit-D and Direct Threat, which is part of the Officer Performance in the FOS Family. All regressions in the table include stratum fixed effects and additional officer-level covariates (race, gender, experience; as well as baseline values of discretionary arrests, officer injuries and an index of officer activities, key outcomes from the administrative data measured similarly at baseline and endline). *** is significant at the 1% level, ** at the 5% level, and * at the 10% level.

Table 3: Key Outcomes in The Field

| | **CM** (1) | **Sit-D** (2) | **SE** (3) | **p-value** (4) | **q-value** (5) |
|---|---|---|---|---|---|
| Uses of non-lethal force | 38.119 | -8.872 | 4.551 | 0.051* | 0.055* |
| Discretionary arrests | 36.849 | -8.474 | 4.292 | 0.048** | 0.055* |
| Officer injuries (days off) | 1.179 | -0.572 | 0.175 | 0.001*** | 0.003*** |
| Officer activities (index) | - | 0.027 | 0.020 | 0.172 | 0.094* |

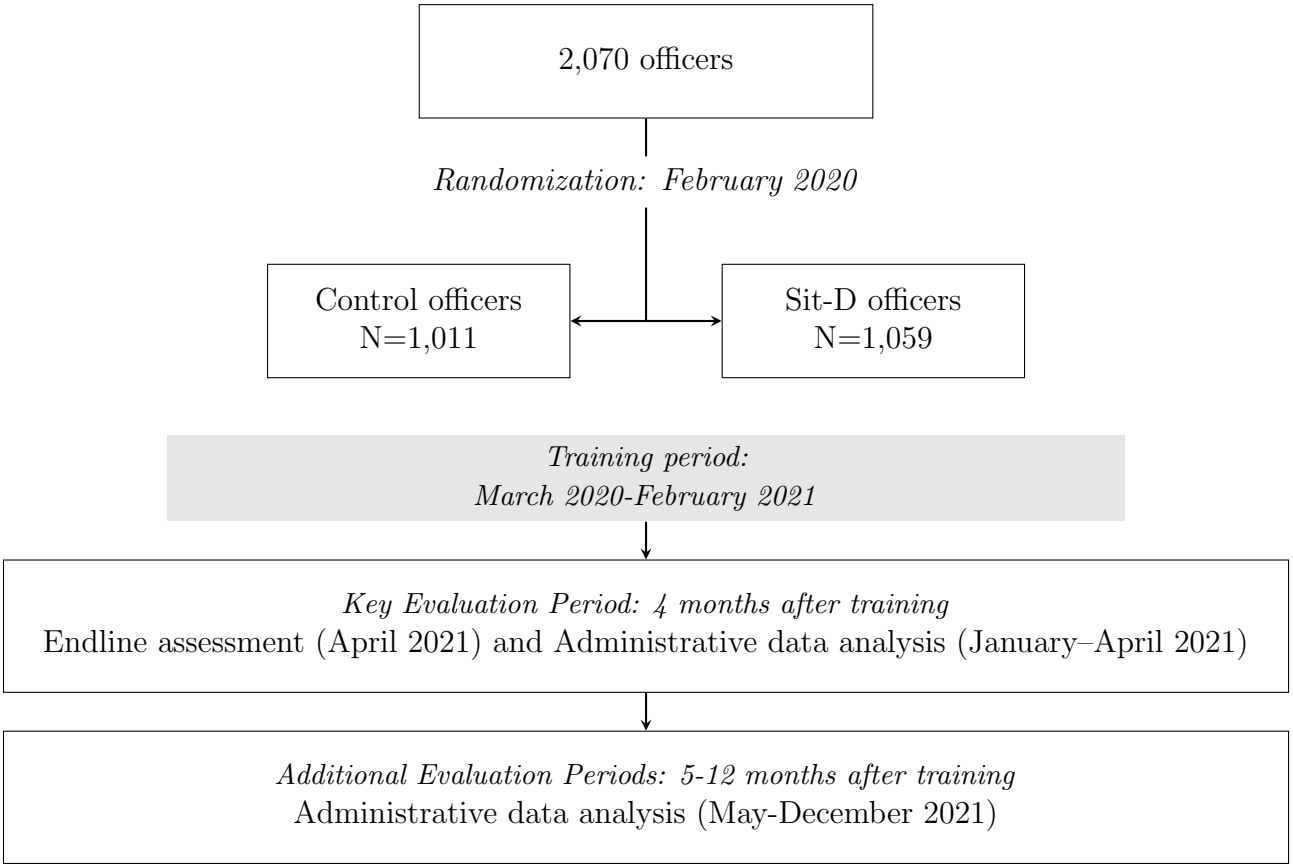
**Notes.** This table shows the effect of Sit-D training on key field outcomes based on estimating equation (2). Each row is a separate regression. Four monthly post-training observations are included for each officer. N=8,070. All regressions include stratum fixed effects, month fixed effects, and additional officer-level covariates (race, gender, experience; as well as baseline values of discretionary arrests, officer injuries and an index of officer activities, key outcomes from the administrative data measured similarly at baseline and endline). Outcomes are measured per 1,000 officers per month, except officer injuries, which is measured per officer per month. Column (1) shows the control mean for each outcome (blank for mean effect indices). Column (2) presents the coefficient on the Sit-D indicator from estimating equation (2). Column (3) shows the standard errors, clustered on officer, and column (4) shows the observed p-value. Column (5) presents the multiple-inference corrected q-values that adjust for the false discovery rate across outcomes in a family. Uses of non-lethal force and discretionary arrests constitute the Adverse Policing Outcomes Family, and officer injuries and the officer activities index constitute the Officer Safety and Activity Family. *** is significant at the 1% level , ** is significant at the 5% level, and * is significant at the 10% level.

Table 4: Arrests of Black Subjects and Other Subjects

| | **CM** (1) | **Sit-D** (2) | **SE** (3) | **p-value** (4) |
|---|---|---|---|---|
| Discretionary arrests: Black subjects | 31.258 | -8.844 | 3.966 | 0.026** |
| Discretionary arrests: Other subjects | 5.591 | 0.371 | 1.634 | 0.821 |
| All arrests: Black subjects | 2016.773 | -219.713 | 103.279 | 0.034** |
| All arrests: Other subjects | 605.083 | -19.949 | 38.627 | 0.606 |
| Other arrests: Black subjects | 1985.515 | -210.869 | 102.155 | 0.039** |
| Other arrests: Other subjects | 599.492 | -20.320 | 38.483 | 0.598 |

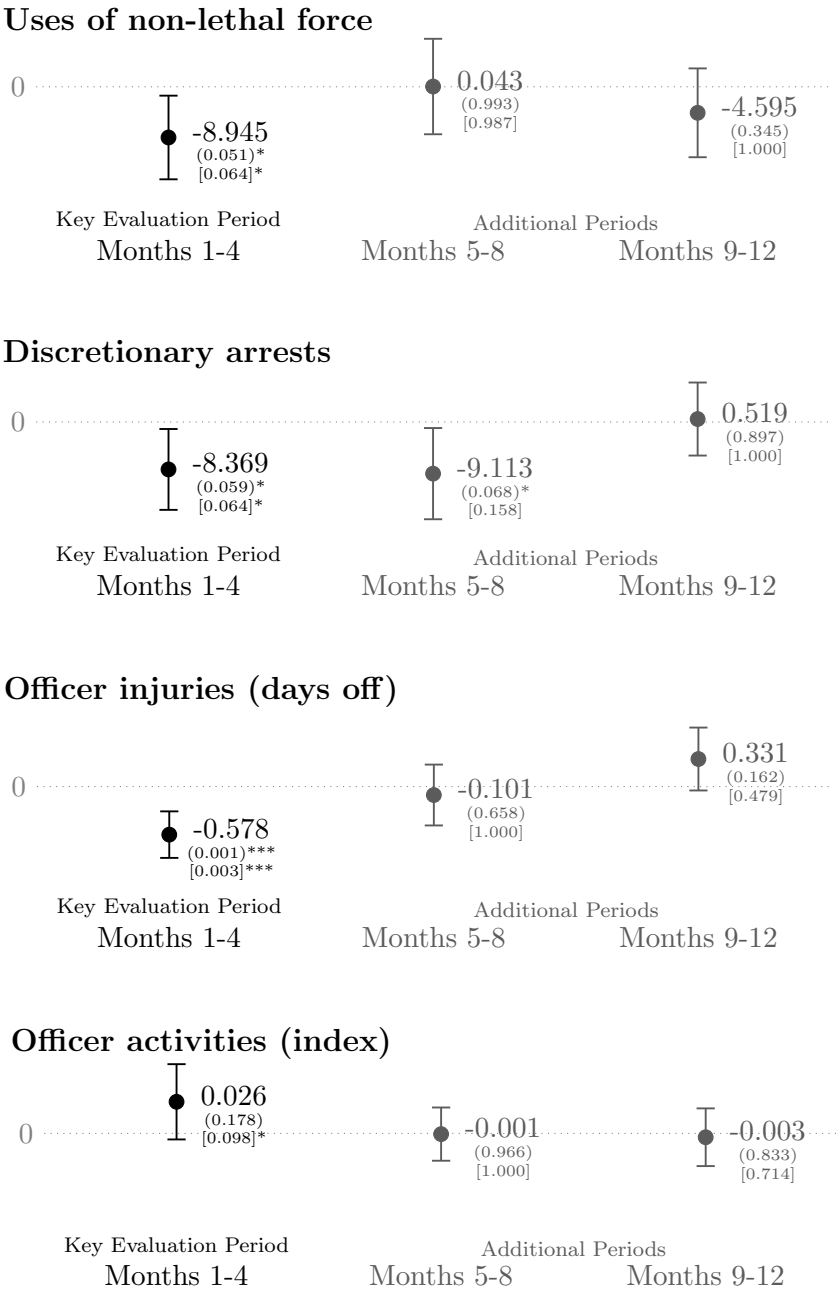
**Notes.** This table shows heterogeneous effects of the Sit-D training on arrests, for Black subjects and subjects of all other races. Each row is a separate regression, based on estimating equation (2). Four monthly post-training observations are included for each officer. N=8,070. Outcomes are measured per 1,000 officers per month. "Discretionary arrests" comprise our pre-specified categories; "other arrests" comprise all other arrests that do not fall under the discretionary arrests variable; and "all arrests" comprise the sum of discretionary and other arrests, spanning all arrests made in that month. All regressions include stratum fixed effects, month fixed effects, and additional officer-level covariates (race, gender, experience; as well as baseline values of discretionary arrests, officer injuries and an index of officer activities, key outcomes from the administrative data measured similarly at baseline and endline). Column (1) shows the control mean for each outcome. Column (2) presents the coefficient on the Sit-D indicator from estimating equation (2). Column (3) shows the standard errors, clustered on officer, and column (4) shows the observed p-value. *** is significant at the 1% level, ** is significant at the 5% level, and * is significant at the 10% level.

Figure 1: Consort Diagram for the Randomized Controlled Trial




**Notes.** This figure shows the months over which randomization and training were conducted. Since officers completed their training at different times, dates for the key evaluation period and additional evaluation periods are shown for the typical officer in the training, who completed the course in December 2020.

Figure 2: Outcomes over Additional Periods




**Notes.** This figure examines the effect of Sit-D training in the key evaluation period and two additional periods, by presenting estimates of equation (3). Each panel is a different regression. Twelve monthly post-training observations are included for each officer. N=23,796. All regressions include stratum fixed effects, month fixed effects, and additional officer-level covariates (see notes to Table 3). The plots show coefficients on the interaction of Sit-D with period indicators along with 90% Confidence Intervals. Observed p-values, based on standard errors clustered on officer, are in parentheses. Multiple-inference corrected q-values that adjust for the false discovery rate within each four-month period and across outcomes in a family are in square brackets. Uses of non-lethal force and discretionary arrests constitute the Adverse Policing Outcomes Family, and officer injuries and the officer activities index constitute the Officer Safety and Activity Family. ***is significant at the 1% level, ** is significant at the 5% level, and * is significant at the 10% level.

# ONLINE APPENDIX

# A Cognitive View of Policing

Oeindrila Dube, Sandy Jo MacArthur & Anuj K. Shah

## Appendix A: Methods

### A.1 Additional Details on Endline Assessment Tasks

Here, we provide additional details on components of the endline assessment described in the Data section of the main paper.

**Knowledge of Sit-D Concepts and Self-Regulation Questions.** The first part of the survey assessed whether officers retained basic knowledge from the training, such as the definitions of different thinking traps (responses are grouped into the "Knowledge of Sit-D Concepts Index"). Next, the survey assessed officers' strategies for regulating stress and emotions ("Coping With Stress Index" and "Emotion Regulation Index").

**Driver's Actions Task.** In this task, officers watched a nine-second video clip in which police stopped a vehicle driven by a Black man. The driver immediately stepped out of the car and ran over to open the rear passenger-side door. Small details in the scene show that the driver stopped near a hospital. Officers spent one minute writing down as many interpretations of the driver's actions as they could think of.

Independent coders counted the total number of reasons the officer listed, whether the officer listed reasons from more than one category (assistance, enforcement, "other"), and whether at least one of the reasons was related to assistance, enforcement, or "other" reasons.

**Pictures Task.** Officers viewed five photos depicting ambiguous situations, where it was unclear if a person in the photos was committing a crime. Each photo included features or clues that could support either a criminal or non-criminal interpretation. One photo showed a person reaching through the window of a car (where they might be breaking into the car or might simply be locked out of their own car). Another photo showed a person using a tool on a window on a house (where they might be breaking into the house or just repairing the window on their own house). A third photo showed a person cutting the lock on a bike (where they might be stealing the bike or trying to break the lock off of their own bike). A fourth photo showed a person spray painting a wall (where they might be tagging it with gang signs or might be legally painting a mural). Another photo showed an altercation in a convenience store (where the person wielding a firearm could have been a robber or a security guard). Three photos were used in the officer-timed version of the task; and two photos were used in the 3-second version of the task.

Note that officers were randomly assigned to see pictures of either a Black or White person in each photo (for each officer, the person's race was consistent across the pictures). Photos were identical except for the person's race. However, it became clear that social desirability would make it difficult to interpret any differences based on the race of the subject depicted. In the control group, officers ascribed criminal intent to 49% of the White subjects but to just 36% of the Black subjects. The lower criminal attribution to Black subjects is consistent with experimenter demand effects, under which participants choose responses in anticipation of what they believe the experimenter wants to hear. As such, our primary analyses focus on responses pooled across Black and White subjects.

**Use of Force Policy.** This task included three videos. One video depicted a person holding a knife while ranting at an officer. Another video depicted an altercation in which one person smashes a bottle on the head of another person. In the third, two-part video, the first part showed a person firing a weapon at someone in a parking lot, and the second part showed the person throwing down their weapon and putting their hands up to surrender to the officer.

Our main analyses here focus on assessing the appropriateness of officers' responses and whether they updated their responses to the two-part video. However, we also developed indices of whether the officer characterized the assailant correctly and specified the force level correctly, as outlined by the Department's Use of Force Policy. These items assess officers' knowledge of department policy.

**Confidence.** The survey also included items to assess whether Sit-D affected officers' confidence ("Confidence Index"). Officers were asked five questions:

- How confident are you in your ability to effectively carry out all aspects of your duty as a police officer?
- How confident are you in your ability to effectively respond to a domestic disturbance call?
- How confident are you in your ability to effectively respond to a robbery in progress call?
- How confident are you in your ability to effectively respond to a shots fired call?
- How confident are you in your ability to do your job effectively during a protest about policing?

**Personalization.** In this task, officers listened to audio recordings of actual officer-civilian interactions and then rated how much they thought the civilian was trying to antagonize the officer ("Personalization Index"). One audio clip depicted a person telling an officer that they have the right to film them while the officer performs a traffic stop. The second depicted a group of people demanding to know an officer's badge number. The third clip

depicted a subject refusing to show their ID to an officer. The fourth clip depicted a person swearing at an officer as the officer initiates a search. And the fifth clip depicted a large crowd of protesters swearing at an officer and chanting about defunding the police.

**FOS Scenarios.** All officers completed three FOS scenarios. Scenario 1 was identical for all officers: It involved responding to a call about a home invasion. Upon arriving at the scene, officers saw the homeowner run out with a gun, and then the armed intruder appears in the scene and opens fire. Officers were randomly assigned to one of two possibilities for Scenario 2: "Husband and Wife" or "Taggers." "Husband and Wife" involved a man holding his wife hostage by pointing a gun at her head, with a crying infant in the background. "Taggers" involved two teens spray painting a wall, with one teen refusing to show their hands to the officer. Officers were also randomly assigned to one of two versions of Scenario 3. Both versions involved an identical street stop, but in Version 1, the subject pulled a cell phone from his pocket and in Version 2, the subject pulled out a handgun from his pocket and fired on the officer.

As specified in our PAP, we also draw on these scenarios to measure the extent to which officers shoot at those who present direct threats (intruder in Scenario 1; man who pulls out a gun in Scenario 3) versus those who do not (homeowner in Scenario 1; man holding wife hostage in Scenario 2).[27]

Officers also answered two questions to assess recall of details from the street stop scenario, which was the final scenario they completed ("Recall Index"). Finally, officers were asked to articulate (i) what actions they took in this scenario and (ii) why they took these actions. Officers' responses were scored on a scale from 1-10 in terms of quality ("Articulation Index"). These latter measures have notable shortcomings. Officers often mentioned that some of the details asked about in the recall questions were irrelevant to how they would respond (e.g., the color of the subject's shoes). And both treatment and control officers wrote

[27]In keeping with our PAP, "Taggers" is not included in this analysis because it was not sufficiently ambiguous to officers and therefore generated little variation in officer decisions to shoot (i.e., few if any officers would shoot in this situation).

very little in response to the articulation questions, which may have been due to fatigue as these were the final questions of the endline assessment.

## A.2 Additional Outcomes from CPD's Administrative Data

**Details on Uses of Force and other TRR Outcomes.** The Tactical Response Reports (TRRs) we use to measure uses of force incidents are divided into 3 categories in our post-training period. Level 1 incidents include actions such as pressure point compliance, joint manipulation, wristlocks, armbars, leg sweeps, weaponless defense techniques, and takedowns that *do not* result in injury. Level 2 incidents include more serious forms of force such as leg sweeps, takedowns, stunning techniques or weaponless direct mechanical actions that *do* result in injury; as well as impact weapon strikes, OC Spray, TASER, canines, impact munitions, force against a handcuffed subject, accidental firearms discharge, and use of firearms to deter an animal. Level 3 incidents include incidents such as discharging a firearm (outside of discharge that is accidental or aimed at deterring an animal), as well as using a choke-hold or an impact weapon to strike someone in the head.

The TRRs also contain additional information on subject injury and tactics, which we present as auxiliary analyses in the appendix given measurement challenges inherent in these variables. Regarding subject injury, officers record whether they believe subjects were injured, and subjects can also allege if they were injured. However, these data are noisy. Only 13% (16%) of TRRs are associated with officer-recorded (subject-alleged) injury, yielding small samples. Moreover, these two measures often do not line up. This can occur for a number of reasons. For example, a subject may experience a minor injury which they do not consider serious enough to allege as an injury. Conversely, officers may miss a potential injury if they interview subjects at an earlier period before an injury is apparent.

In addition, the TRRs report on subject hospitalizations. However, these data are subject to another measurement challenge: Many hospitalizations do not arise from injury or the use of force itself, but stem from the subject's drug-use, mental-health issues, or other pre-

existing health conditions. To attempt to focus on hospitalizations that correspond to subject injury, we also created an additional indicator for whether the subject was both hospitalized *and* either the subject alleged injury or the officer recorded an injury. This is nonetheless an imperfect proxy for injury arising from the officer's actions.

Finally, TRRs contain information on officers' tactics in use of force incidents. We used this to create an index designed to measure if officers relied less on force tactics (strikes, kicks, take-downs, TASERs, and other reportable forces), and if they relied more on other types of tactics (such as giving verbal direction, movement to avoid attack, tactical positioning, and establishing a zone of safety). We also view this as an auxiliary measure, since we can only observe these tactics conditional on an officer entering into a use of force incident (i.e., we cannot observe if they used certain tactics to avoid entering into a force incident in the first place). Our main measure, the number of force incidents, better reflects the effort officers may have expended to avoid using force.

**Downstream Outcomes.** We also examine two additional outcomes that are downstream responses to an officer's actions. We obtain information on complaints levied against officers from CPD's complaint management system. We use this to measure total complaints and to create an index of categories associated with force and abuse, which includes accusations related to excessive force, civil rights violations, verbal abuse, arrest/lockup, domestic violence, conduct unbecoming (for altercations, disturbances, and harassment), as well as operation personnel violations (for categories such as neglect of duty and inadequate response). This index is fairly comprehensive, spanning complaint categories used in prior work (Ba, 2020).

While many complaints originate from community members, approximately 15% are generated internally from other CPD members. We are not able to distinguish where the complaint originates in our data. It is important to note that the complaint data are incomplete and thus potentially noisy, since it takes 5 months on average for an incident to work its way through the system and enter into the administrative data.

Finally, we use data from CPD's personnel performance system to measure awards and commendations. We use this to create an index of honorable mentions, department commendations, and other high-level awards allocated to individual active-duty officers. While these awards may contain some signal, political and bureaucratic considerations in allocating awards make them potentially noisy measures of officer performance.

Table A1: Select Sit-D Activities

| Activity Category | Examples | Purpose |
|---|---|---|
| **Icebreakers:** Officers watch short videos that contain subtle scene changes, and they try to identify all of the changes. These are often done at the start of class or when coming back from a break to re-engage officers. | **Invisible Gorilla:** The video shows people passing a basketball around. Officers count the number of passes. Meanwile, the scene undergoes many changes: a person dressed in a gorilla suit walks through the scene, the curtains change colors, basketball players enter and exit the scene.<br>**Whodunnit:** The video appears to show a murder mystery drama. But details in the scene keep changing (e.g., props are added and removed). | To get officers actively participating in discussions, with an emphasis on processing information more deliberately (e.g., noticing visual details they might have initially overlooked). |
| **Subjective/Objective Discussions:** Officers engage in a mix of activities that highlight the distinction between their subjective impressions and objective facts. There are several of these exercises, particularly near the beginning of the training. | **Picture Communication:** Officers work with a partner. One officer describes a painting that they see to their partner (who cannot see the painting). The partner then tries to identify the painting in a lineup of similar paintings. Officers then discuss the details they subjectively assumed were important for communication, and which features turn out to actually be important for communication.<br>**Camera View:** Officers discuss "what a camera would see." In these exercises, they are not permitted to make subjective statements (e.g., guesses about a person's intent). They can only describe the objective facts of a situation. They then discuss how those facts could be interpreted differently. | To highlight for officers how their initial subjective impressions of a situation can make it hard for them to notice some objective facts. This sets the stage for the importance of considering alternative interpretations. |
| **Breathing Exercises:** Officers learn a variety of breathing techniques. Some breathing exercises are done without audiovisual stimuli, and others are done while listening to radio calls or watching police scenes play out. There are 18 of these exercises spread throughout the 4 sessions. | **Shots Fired Radio Call:** Officers practice breathing exercises while listening to a chaotic radio call that includes shots fired and injured officers. The call is taken from an evening when officers were ambushed and shot in Dallas in 2016.<br>**Foot Pursuit Video:** Officers practice breathing exercises while watching surveillance footage of a foot pursuit through housing projects that results in shots fired. | To help officers remain calm during stressful situations so that they can then deliberately think through alternative interpretations. |

Select Sit-D Activities (continued)

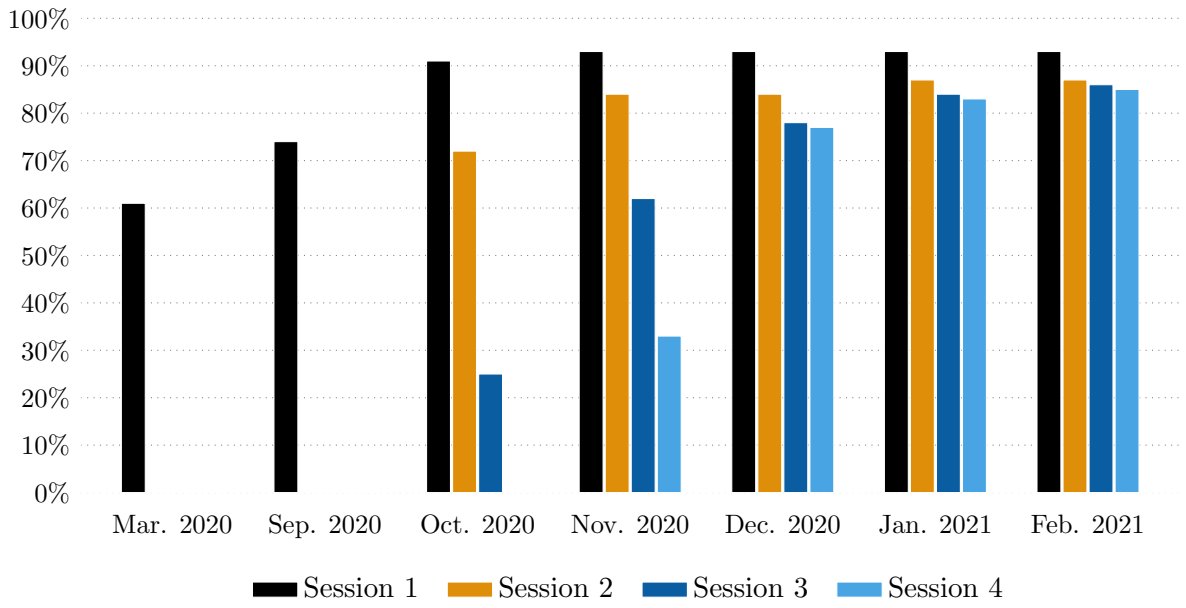
| Activity Category | Examples | Purpose |
|---|---|---|
| **Radio Calls:** Officers work with a partner in the classroom while listening to a recorded radio call for service. Officers prepare as if they are en route to the call, working through a checklist that prompts them to discuss resources they will need, scenarios they might encounter, information they can gather, and different courses of action they might take. There are 7 of these exercises throughout the 4 sessions. | **Robbery Suspect:** Officers prepare to join the search for a robbery suspect, with the dispatcher and pursuing officers adding more information as the call unfolds.<br>**Police Protest:** Officers prepare to arrive at the scene of a large protest that is moving through city streets, with the dispatcher and officers on the scene discussing crowd control strategies. | To help officers consider alternative interpretations of situations prior to arriving on scene. |
| **Video Vignettes:** These videos consist of a mix of real bodycam footage and filmed scenarios (based on real cases), chosen because they depict ambiguous situations with numerous plausible interpretations. Officers watch these videos and then privately write down different interpretations of the situation. Officers then discuss as a group, highlighting cues they may have missed and interpretations they may have overlooked. Officers also discuss how different thinking traps might have shaped their responses. There are 10 of these exercises throughout the 4 sessions. | **Young Woman at Apartment:** In this video, a neighbor flags down an officer and tells him there is a young woman pacing in front of her apartment in bare feet (on a snowy day). The officer questions the young woman, and she replies with one-word, tentative answers. At that point a man comes out and explains to the officer that the young woman is his girlfriend's sister whom he's taking care of, and he takes the young woman back inside, at which point the video stops and officers in the training discuss what might be going on. This is based on a case in which the young woman was kidnapped and sexually assaulted, but felt unable to communicate with a male officer given the trauma she experienced.<br>**Man Near Dumpster:** In this video, officers respond to a radio call about a man near a dumpster who is acting in a threatening manner. As the officers arrive on scene, they see that the man appears to be shouting at someone out of view, threatening them with a bottle in his hand. The man does not respond to the officers. At this point, the video stops and officers in the training discuss what might be going on. This is based on a case in which the man was schizophrenic and was not threatening another person, but was in crisis himself. | To highlight how multiple officers might see a situation differently, underscoring the importance of considering alternative interpretations and searching for information that can help identify the most accurate interpretation. |

Select Sit-D Activities (continued)

| Activity Category | Examples | Purpose |
|---|---|---|
| **Force Options Simulations:** Officers work through simulations in which they interact with life-sized projections of subjects. Officers can use retrofitted equipment (e.g., firearms, TASERs, OC spray), while trainers control how subjects respond. Officers then participate in active debriefs of the simulations in which they discuss their interpretation of the situation and the reasoning behind their action. These debriefs push officers to consider cues they may have overlooked, along with alternative interpretations and courses of action. | **Dumpster Divers:** Officers see a person diving in a dumpster outside of an industrial building. While questioning the first suspect, another suspect suddenly emerges from the dumpster as well. Eventually the second suspect reaches into their coat pocket and quickly pulls something out. Some officers see the person pull out a weapon, while others see them pull out a screwdriver and then put their hands up.<br>**Suicidal Man:** Officers respond to a call for a man having a mental health crisis. The man is holding a large knife in a parking lot. There is a healthcare professional nearby pleading with the person to drop their weapon and not to harm themselves. Officers primarily interact with the man in crisis. In some scenarios, the man suddenly attacks the healthcare professional with the knife, while in other scenarios the man either drops the knife or stabs himself, depending on how officers interact with him. | To practice the Thinking Tactic Model during realistic scenarios so that officers are better prepared to regulate their emotions and stress while considering alternative interpretations in the field. Also, to help officers recognize how the force options they employ are tied to their interpretation, and how different interpretations might suggest using different force options. |

Figure A1: Training Content and Participation Over Time

| Session | Concepts Introduced |
|---|---|
| **Session 1** | Importance of alternative interpretations<br>Stress response and breathing tactics<br>Thinking Tactic Model<br>Triggers<br>Catastrophizing and minimizing<br>Confirmation trap<br>Pre-scenario checklist |
| **Session 2** | Overgeneralization<br>Personalization<br>All-or-none thinking<br>Anchoring<br>Thinking traps and duty to intervene |
| **Session 3** & **Session 4** | Applying concepts in exercises and FOS scenarios |




**Notes.** The top panel shows when Sit-D concepts are introduced across the four sessions. These concepts are taught through a variety of exercises (see Table A1). No new concepts are introduced in Sessions 3 and 4. Instead, trainees focus on applying those concepts in additional exercises and FOS scenarios. The bottom panel shows the cumulative proportion of the 1,059 officers assigned to take the Sit-D training over time. No more than 5 control officers showed up to any training sessions, and no control officers attended enough sessions to be considered trained.

Table A2: Discretionary Arrest Statutes

| **Statute** | **Description** |
|---|---|
| 8-4-010(A) | Disorderly conduct - breach of peace |
| 8-4-010(B) | Disorderly conduct - offensive act or gesture |
| 8-4-010(C) | Disorderly conduct - failure to cease conduct |
| 8-4-010(D) | Disorderly conduct - failure to obey order to disperse |
| 8-4-010(E) | Disorderly conduct - failure to obey police |
| 8-4-010(G) | Disorderly conduct - blocking access to commercial establishment |
| 9-88-010 | Refusing to comply with order from a police officer, firefighter, or person directing traffic |
| 510 ILCS 68.0/105-45 | Obstructing an officer |
| 515 ILCS 5.0/1-200 | Obstructing an officer |
| 520 ILCS 5.0/1.22 | Resisting or obstructing an officer |
| 625 ILCS 40.0/2-4 | Resisting or obstructing an officer |
| 625 ILCS 5.0/11-203 | Refusing to comply with order from a police officer, firefighter, or person directing traffic |
| 625 ILCS 5.0/18B-103.1-A | Refusing to comply with order from an officer |
| 720 ILCS 5.0/26-1.1-A | Disorderly conduct - false report to defraud an insurer |
| 720 ILCS 5.0/26-1-A-1 | Disorderly conduct - breach of peace |
| 720 ILCS 5.0/26-1-A-2 | Disorderly conduct - false fire alarm |
| 720 ILCS 5.0/26-1-A-3 | Disorderly conduct - false bomb threat |
| 720 ILCS 5.0/26-1-A-4 | Disorderly conduct - false report of an offense |
| 720 ILCS 5.0/26-1-A-9 | Disorderly conduct - false request for an ambulance |
| 720 ILCS 5.0/31-1-A | Resisting or obstructing a peace officer, firefighter, or correctional institution employee |
| 720 ILCS 5.0/31-1-A-7 | Resisting or obstructing a peace officer, firefighter, or correctional institution employee, and causing injury |
| 720 ILCS 5.0/31-4.5-A | Obstructing identification |

**Notes.** This table lists the statutes included in our measure of discretionary arrests. Statutes beginning with 8 or 9 are from the Municipal Code of Chicago. Statutes beginning with 510 are from the Illinois statutes concerning animals. Statutes beginning with 515 are from the Illinois statutes concerning fish and aquatic life. Statutes beginning with 520 are from the Illinois statutes concerning wildlife. Statutes beginning with 625 are from the Illinois vehicle code. Statutes beginning with 720 are from the Illinois Criminal Code.

Table A3: Families of Outcomes

| Family | Outcomes |
|---|---|
| **Knowledge** | Knowledge Of Sit-D Concepts Index |
| | Correct assailant level in policy (index) |
| | Correct force level in policy (index) |
| | Characterization of assailant who is a direct threat (z-score) |
| **Navigating Cognitively Demanding Situations** | Coping With Stress Index |
| | Emotional Regulation Index |
| | Total explanations |
| | Explanations from multiple categories |
| | At least one explanation - assistance category |
| | At least one explanation - enforcement category |
| | At least one explanation - other category |
| | Alternative Features Index (both tasks) |
| | Confirming Features Index (both tasks) |
| | Criminal Interpretations Index (both tasks) |
| | Index of decision time (both tasks) |
| | Index of processing time (officer-timed task) |
| | Change in perceived threat and force assessment (index) |
| | Appropriate actions (index) |
| | Inappropriate actions (index) |
| | Confidence Index |
| | Personalization Index |
| **Performance in the FOS** | Did the officer communicate with the person? (index) |
| | Did the officer give verbal direction/ commands to the person? (index) |
| | Did the officer radio dispatch? (index) |
| | Did the officer freeze during the scenario? (index) |
| | Did the officer kneel or move to cover/ concealment? (index) |
| | Shooting in the FOS (interaction term) |
| | Recall Index |
| | Articulation Index |
| **Adverse Policing Outcomes** | Use of non-lethal force |
| | Discretionary arrests |
| **Officer Safety and Activity Family** | Officer injuries (days off) |
| | Officer activities (index) |
| **Auxiliary TRR** | Subject injuries (officer reported) |
| | Subject allegations of injures |
| | Hospitalization |
| | Hospitalizations and either subject alleged injury or officer reported an injury |
| | Tactics used in use of force incidents (index) |
| **Downstream Actions From Officers' Actions** | Commendations and awards |
| | Total complaints |
| | Force and abuse related complaints (index) |

**Notes.** This table lists how conceptually related indices and outcomes are grouped together into broad families which are used for the purposes of adjusting inference for multiple hypothesis testing.

# Appendix B: Additional Analysis and Results

## Additional Analysis Tables



## Additional Analysis Figures

Table B1: Balance on Key Covariates (Full Sample)

| | Control Mean | Treatment Mean | Difference | N |
|---|---|---|---|---|
| **Panel A: Officer characteristics** | | | | |
| Age | 37.840<br>(8.512) | 38.052<br>(8.653) | 0.190<br>(0.333) | 2,070 |
| Years of experience | 9.229<br>(7.397) | 9.219<br>(7.208) | -0.019<br>(0.277) | 2,070 |
| Gender: Male | 0.799<br>(0.401) | 0.813<br>(0.390) | 0.015<br>(0.017) | 2,070 |
| Race and ethnicity: Black | 0.132<br>(0.338) | 0.125<br>(0.330) | -0.010<br>(0.014) | 2,070 |
| Race and ethnicity: Hispanic | 0.363<br>(0.481) | 0.342<br>(0.475) | -0.019<br>(0.021) | 2,070 |
| Race and ethnicity: White | 0.449<br>(0.498) | 0.476<br>(0.500) | 0.028<br>(0.021) | 2,070 |
| Race and ethnicity: Other | 0.056<br>(0.231) | 0.058<br>(0.233) | 0.001<br>(0.010) | 2,070 |
| **Panel B: Officer performance prior to treatment** | | | | |
| Uses of force | 1.695<br>(2.535) | 1.600<br>(2.354) | -0.090<br>(0.098) | 2,070 |
| Uses of force - All but lethal | 1.672<br>(2.499) | 1.568<br>(2.309) | -0.098<br>(0.096) | 2,070 |
| Subject injuries (officer reported) | 0.213<br>(0.626) | 0.216<br>(0.586) | 0.003<br>(0.026) | 2,070 |
| Subject allegation of injuries | 0.179<br>(0.514) | 0.185<br>(0.518) | 0.006<br>(0.022) | 2,070 |
| Officer injuries (days off) | 14.528<br>(44.546) | 14.418<br>(46.310) | -0.178<br>(1.980) | 2,070 |
| Tactics used in TRRs (index) | 0.000<br>(0.348) | -0.018<br>(0.278) | -0.018<br>(0.014) | 2,070 |
| Discretionary arrests | 3.581<br>(4.299) | 3.665<br>(4.489) | 0.105<br>(0.152) | 2,070 |
| Total officer activities (index) | 0.000<br>(0.405) | 0.017<br>(0.432) | 0.019<br>(0.016) | 2,070 |
| Complaints | 1.180<br>(1.705) | 1.091<br>(1.621) | -0.085<br>(0.068) | 2,070 |
| Hospitalizations | 0.555<br>(0.964) | 0.520<br>(0.987) | -0.034<br>(0.041) | 2,070 |
| Awards and commendations | 16.616<br>(16.657) | 17.008<br>(17.709) | 0.517<br>(0.588) | 2,070 |
| F-test: p-value = 0.3939 | | | | 2,070 |

**Notes.** This table examines baseline balance in officer characteristics and officer outcomes in CPD's administrative data during the two years preceding randomization (over January 2018 – January 2020). The first column shows the mean of the control group at baseline; the second column shows the mean of the treatment group at baseline; and the third column presents the difference in means between the treatment and control groups. These estimates are attained by regressing each covariate on the Sit-D indicator, along with stratum (unit x watch) fixed effects. The last column shows the number of observations in these regressions. The last row of the table presents the p-value associated with an F-test of joint significance, from a regression of the Sit-D treatment indicator on all the variables examined in the table, along with stratum fixed effects. *** p <0.01, ** p <0.05, * p <0.1.

Table B2: Balance on Key Covariates (Endline Assessment Sample)

| | Control Mean | Treatment Mean | Difference | N |
|---|---|---|---|---|
| **Panel A: Officer characteristics** | | | | |
| Age | 37.712 | 38.011 | 0.432 | 1,696 |
| | (8.165) | (8.560) | (0.364) | |
| Years of experience | 9.173 | 9.204 | 0.186 | 1,696 |
| | (7.225) | (7.114) | (0.306) | |
| Gender: Male | 0.795 | 0.820 | 0.024 | 1,696 |
| | (0.404) | (0.384) | (0.019) | |
| Race and ethnicity: Black | 0.126 | 0.126 | -0.001 | 1,696 |
| | (0.332) | (0.332) | (0.015) | |
| Race and ethnicity: Hispanic | 0.370 | 0.343 | -0.024 | 1,696 |
| | (0.483) | (0.475) | (0.023) | |
| Race and ethnicity: White | 0.448 | 0.472 | 0.024 | 1,696 |
| | (0.498) | (0.500) | (0.024) | |
| Race and ethnicity: Other | 0.056 | 0.058 | 0.001 | 1,696 |
| | (0.230) | (0.235) | (0.011) | |
| **Panel B: Officer performance prior to treatment** | | | | |
| Uses of force | 1.596 | 1.578 | -0.054 | 1,696 |
| | (2.311) | (2.327) | (0.103) | |
| Uses of force - All but lethal | 1.573 | 1.546 | -0.061 | 1,696 |
| | (2.276) | (2.274) | (0.101) | |
| Subject injuries (officer reported) | 0.199 | 0.206 | 0.002 | 1,696 |
| | (0.577) | (0.553) | (0.027) | |
| Subject allegation of injuries | 0.162 | 0.179 | 0.011 | 1,696 |
| | (0.488) | (0.509) | (0.023) | |
| Officer injuries (days off) | 13.167 | 13.001 | -0.671 | 1,696 |
| | (42.983) | (44.358) | (2.104) | |
| Tactics used in TRRs (index) | 0.000 | -0.003 | -0.003 | 1,696 |
| | (0.082) | (0.073) | (0.004) | |
| Discretionary arrests | 3.475 | 3.743 | 0.183 | 1,696 |
| | (4.198) | (4.615) | (0.170) | |
| Total officer activities (index) | 0.000 | 0.016 | 0.017 | 1,696 |
| | (0.235) | (0.291) | (0.012) | |
| Complaints | 1.077 | 1.088 | 0.002 | 1,696 |
| | (1.529) | (1.669) | (0.073) | |
| Hospitalizations | 0.529 | 0.506 | -0.031 | 1,696 |
| | (0.910) | (0.960) | (0.043) | |
| Awards and commendations | 16.165 | 17.375 | 0.933 | 1,696 |
| | (14.563) | (18.075) | (0.598) | |
| F-test: p-value = 0.5733 | | | | 1,696 |

**Notes.** This table examines baseline balance in officer characteristics and officer outcomes, restricting the sample to officers who completed an endline assessment. The first column shows the mean of the control group at baseline; the second column shows the mean of the treatment group at baseline; and the third column presents the difference in means between the treatment and control groups. These estimates are attained by regressing each covariate on the Sit-D indicator, along with stratum (unit x watch) fixed effects. The last column shows the number of observations in these regressions. The last row of the table presents the p-value associated with an F-test of joint significance, from a regression of the Sit-D treatment indicator on all the variables examined in the table, along with stratum fixed effects. *** p <0.01, ** p <0.05, * p <0.1.

Table B3: Attrition

| | **Control Mean** | **Treatment Mean** | **Difference** | **N** |
|---|---|---|---|---|
| Attrition | 0.087 | 0.094 | 0.008<br>(0.013) | 2,070 |
| Attrition (12 months) | 0.010 | 0.011 | 0.001<br>(0.004) | 2,040 |
| Attrition (8 months) | 0.018 | 0.026 | 0.008<br>(0.006) | 2,044 |
| Attrition (4 months) | 0.045 | 0.043 | -0.002<br>(0.009) | 2,052 |
| Attrition (Endline survey) | 0.169 | 0.192 | 0.022<br>(0.017) | 2,070 |

**Notes.** This table examines whether attrition is predicted by treatment status. The first four rows measure attrition out of CPD's administrative data, gauging if the officer is missing in this data source for all months after January 2021 (in the top row); over March 2021-February 2022 (in the second row); over July 2021-February 2022 (in the third row); and over November 2021 – February 2022 (in the bottom row). The bottom row defines attrition as missing from the endline assessment. Each row represents a different regression in which the attrition indicator is regressed on the Sit-D indicator. All regressions include stratum (unit x watch) fixed effects. Robust standard errors are shown in parentheses. *** p <0.01, ** p <0.05, * p <0.1.

Table B4: Unit Switching

| | **Control Mean** (1) | **Sit-D** (2) | **SE** (3) | **N** (4) |
|---|---|---|---|---|
| **Panel A: Unit × Watch Switch** | | | | |
| Ever switched in any post-period month | 0.460 | -0.010 | (0.020) | 1,996 |
| Switched for more than half the months in post-period | 0.393 | -0.012 | (0.020) | 1,996 |
| Switched for all months in post-period | 0.297 | 0.015 | (0.018) | 1,996 |
| **Panel B: Unit Switch** | | | | |
| Ever switched in any post-period month | 0.328 | 0.010 | (0.018) | 1,996 |
| Switched for more than half the months in post-period | 0.274 | 0.009 | (0.017) | 1,996 |
| Switched for all months in post-period | 0.212 | 0.019 | (0.016) | 1,996 |

**Notes.** This table examines if the Sit-D treatment affected officer tendencies to switch away from the location in which they were working at the time of randomization. Panel A measures switching away from the unit × watch; and Panel B from just the unit. Each panel presents three measures: whether the officer ever switched in any post-training month; switched for more than half the post-training months; or switched for all of the post-training months. Each row is one regression. All regressions include stratum fixed effects. Column (1) shows the control means for each outcome. Column (2) presents the coefficients on the Sit-D indicator. Column (3) shows robust standard errors in parentheses. Column (4) shows the number of observations in each regression. *** is significant at the 1% level, ** is significant at the 5% level, and * significant at the 10% level.

## B.1 Additional Endline Results

In this section, we describe additional results from the endline assessment related to which concepts officers recalled from the training, which stress and emotion-regulation strategies they report using, how confident they feel, and the extent to which they might personalize interactions. We also discuss additional results on what officers recalled and articulated after the simulator exercises (see Appendix A.1 for more details on the procedures).

**Knowledge of Training Concepts and Use of Force Policy.** In Table B5, we find that Sit-D officers recalled significantly more core constructs from the training, nearly .6 SDs above control officers. The results suggest that the course was delivered well and officers retained key lessons from the training, even four months after the classes wrapped up. In contrast to these effects, we do not see strong evidence that the training changes knowledge of CPD's Use of Force Policy in Table B6, as the coefficients on the measures show varied signs and none of the effects here remain significant after multiple-inference correction. Importantly, Sit-D is not a training on department policies and it does not explicitly instruct officers on the Use of Force Policy, which is the subject of another mandatory training required of all officers.

**Self-Regulation Strategies.** The first two steps of the Thinking Tactic Model focus on recognizing emotional triggers and using self-regulation strategies to lay the groundwork for greater deliberation. Table B7 shows that Sit-D affected officers' strategies for coping with stress and regulating emotions, as measured by their respective indices. When we examine the components of these indices, we observe strong evidence that Sit-D officers are more likely to use various strategies to cope with stress, including deep breathing (a point of emphasis in the training). Perhaps more striking, we also see that Sit-D officers use additional strategies to regulate their emotions. In fact, Sit-D officers are more likely to control their emotions by changing the way they see the situations they are in. That is, they control their emotions in

part by drawing on our key mechanism: considering alternative interpretations.

**Confidence.** As shown in Table B8, Sit-D officers feel greater confidence in handling their duties in the field. This suggests that the training not only changes how officers think through situations, but also their perceptions of their ability to navigate those situations.

**Personalization.** In Table B9, we examine effects on personalization, or the extent to which officers think subjects are trying to antagonize them. We do not see significant differences between Sit-D and control officers in the Personalization Index. The coefficients on the individual recordings show varied signs and levels of precision.

Of course, we cannot distinguish between limitations of our measurement or whether the training had no effect on the extent to which officers personalize situations. However, one possible shortcoming of our measurement might be that the audio clips officers listened to were stripped of all context. This might have made it more difficult for officers to think of alternative reasons why a subject might be acting a certain way. In more realistic scenarios, it is possible that Sit-D officers would have found ways to de-personalize the situation.

**Performance in the Simulators.** Finally, we do not see significant effects of Sit-D on officers' recall of details or articulation of their actions, which were measured after the FOS exercises ended (see Table B10). This may reflect the shortcomings of these measures noted in Appendix A.1 —namely, the details officers were asked to recall were irrelevant to the scenario and fatigue at the end of the assessment may have limited articulation in treatment and control.

Table B5: Knowledge of Training Concepts

| | **CM** (1) | **Sit-D** (2) | **SE** (3) | **p-value** (4) | **q-value** (5) |
|---|---|---|---|---|---|
| Knowledge Of Sit-D Concepts Index | - | 0.597 | 0.029 | $<0.001^{***}$ | $0.001^{***}$ |
| Confirmation Trap | 0.053 | 0.105 | 0.015 | $<0.001^{***}$ | |
| Personalization | 0.340 | 0.418 | 0.022 | $<0.001^{***}$ | |
| Overgeneralization | 0.245 | 0.194 | 0.023 | $<0.001^{***}$ | |
| Catastrophizing | 0.539 | 0.309 | 0.021 | $<0.001^{***}$ | |
| Thinking Tactic Model | 0.649 | 0.165 | 0.013 | $<0.001^{***}$ | |

**Notes.** This table shows the effect of Sit-D training on officers' knowledge of key concepts from the training (measured in the endline assessment), based on estimating equation (1). Each row is a different regression. One observation is included for each officer. N=1,669. All regressions include stratum fixed effects and additional officer-level covariates (race, gender, experience; as well as baseline values of discretionary arrests, officer injuries and an index of officer activities, key outcomes from the administrative data measured similarly at baseline and endline). The top row shows the results for the Knowledge Of Sit-D Concepts Index, while the remaining rows show the results for the components of the index. Column (1) shows the control mean for each outcome (blank for mean effect indices). Column (2) shows the coefficients on the Sit-D indicator. Column (3) shows robust standard errors. Column (4) shows the observed p-values. Column (5) shows the multiple-inference corrected q-value that adjusts for the false discovery rate across outcomes in a family. The Knowledge Of Sit-D Concepts Index is part of the Knowledge Family. *** is significant at the 1% level, ** is significant at the 5% level, and * is significant at the 10% level.

Table B6: Knowledge of Use of Force Policy

| | **Sit-D** (1) | **SE** (2) | **p-value** (3) | **q-value** (4) |
|---|---|---|---|---|
| Correct assailant level in policy (index) | 0.027 | 0.037 | 0.462 | 0.151 |
| Correct force level in policy (index) | -0.066 | 0.040 | 0.098* | 0.110 |
| Characterization of assailant who is a direct threat (z-score) | 0.084 | 0.048 | 0.082* | 0.110 |

**Notes.** This table shows the effect of Sit-D training on officers' knowledge of CPD's use of force policy (measured in the endline assessment), based on estimating equation (1). Each row is a different regression. One observation is included for each officer. N=1,669. All regressions include stratum fixed effects and additional officer-level covariates (race, gender, experience; as well as baseline values of discretionary arrests, officer injuries and an index of officer activities, key outcomes from the administrative data measured similarly at baseline and endline). Column (1) shows the coefficients on the Sit-D indicator. Column (2) shows robust standard errors. Column (3) shows the observed p-values. Column (4) shows the multiple-inference corrected q-values that adjust for the false discovery rate across outcomes in a family. All outcomes in this table are part of the Knowledge Family. *** is significant at the 1% level, ** is significant at the 5% level, and * is significant at the 10% level.

Table B7: Stress and Emotion Regulation

| | **CM** (1) | **Sit-D** (2) | **SE** (3) | **p-value** (4) | **q-value** (5) |
|---|---|---|---|---|---|
| **Panel A: Coping with Stress** | | | | | |
| Coping With Stress Index | - | 0.199 | 0.038 | $<0.001^{***}$ | $0.001^{***}$ |
| In stressful situations, how often do you cope with the stress by engaging in deep breathing? | 3.620 | 0.426 | 0.066 | $<0.001^{***}$ | |
| In stressful situations, how often do you cope with the stress by taking a break from the situation if it is possible to do so? | 3.650 | 0.271 | 0.065 | $<0.001^{***}$ | |
| In stressful situations, how often do you cope with the stress by seeking support from others if it is possible to do so? | 3.168 | 0.155 | 0.072 | $0.032^{**}$ | |
| **Panel B: Emotion Regulation** | | | | | |
| Emotion Regulation Index | - | 0.077 | 0.029 | $0.007^{***}$ | $0.038^{**}$ |
| I could be experiencing some emotion and not be conscious of it until some time later. | 4.458 | 0.023 | 0.057 | 0.688 | |
| I control my emotions by changing the way I think about the situation I'm in. | 4.082 | 0.189 | 0.067 | $0.005^{***}$ | |
| I control my emotions by not expressing them. | 3.648 | 0.114 | 0.070 | 0.105 | |

**Notes.** This table shows the effect of Sit-D training on how officers coped with stress and regulated their emotions (measured in the endline assessment), based on estimating equation (1). Each row is a different regression. One observation is included for each officer. N=1,669. All regressions include stratum fixed effects and additional officer-level covariates (race, gender, experience; as well as baseline values of discretionary arrests, officer injuries and an index of officer activities, key outcomes from the administrative data measured similarly at baseline and endline). Panel A shows the results for the Coping With Stress Index (top row) and its components (remaining rows). Panel B shows the results for the Emotion Regulation Index (top row) and its components (remaining rows). Column (1) shows the control mean for each outcome (blank for mean effect indices). Column (2) shows the coefficients on the Sit-D indicator. Column (3) shows robust standard errors. Column (4) shows the observed p-values. Column (5) shows the multiple-inference corrected q-value that adjusts for the false discovery rate across outcomes in a family. The Coping With Stress Index and Emotion Regulation Index are both part of the Navigating Cognitively Demanding Situations Family. *** is significant at the 1% level, ** is significant at the 5% level, and * is significant at the 10% level.

Table B8: Confidence in Policing

| | **CM** (1) | **Sit-D** (2) | **SE** (3) | **p-value** (4) | **q-value** (5) |
|---|---|---|---|---|---|
| Confidence Index | - | 0.094 | 0.041 | 0.021** | 0.063* |
| How confident are you in your ability to effectively respond to a domestic disturbance call? | 3.580 | 0.091 | 0.026 | <0.001*** | |
| How confident are you in your ability to effectively respond to a robbery in progress call? | 3.596 | 0.049 | 0.026 | 0.054* | |
| How confident are you in your ability to effectively respond to a shots fired call? | 3.614 | 0.040 | 0.026 | 0.127 | |
| How confident are you in your ability to do your job effectively during a protest about policing? | 3.634 | 0.043 | 0.026 | 0.093* | |
| How confident are you in your ability to effectively carry out all aspects of your duty as a police officer? | 3.382 | 0.043 | 0.036 | 0.231 | |

**Notes.** This table shows the effect of Sit-D training on officers' confidence in their ability to respond to different situations (measured in the endline assessment), based on estimating equation (1). Each row is a different regression. One observation is included for each officer. N=1,669. All regressions include stratum fixed effects and additional officer-level covariates (race, gender, experience; as well as baseline values of discretionary arrests, officer injuries and an index of officer activities, key outcomes from the administrative data measured similarly at baseline and endline). The top row shows the results for the Confidence Index, while the remaining rows show the results for the components of the index. Column (1) shows the control mean for each outcome (blank for mean effect indices). Column (2) shows the coefficients on the Sit-D indicator. Column (3) shows robust standard errors. Column (4) shows the observed p-values. Column (5) shows the multiple-inference corrected q-value that adjusts for the false discovery rate across outcomes in a family. The Confidence Index is part of the Navigating Cognitively Demanding Situations Family. *** is significant at the 1% level, ** is significant at the 5% level, and * is significant at the 10% level.

Table B9: Personalization

| | **CM** (1) | **Sit-D** (2) | **SE** (3) | **p-value** (4) | **q-value** (5) |
|---|---|---|---|---|---|
| Personalization Index | - | 0.027 | 0.036 | 0.448 | 0.400 |
| While an officer performs a traffic stop, a bystander starts filming. | 2.830 | 0.077 | 0.046 | 0.098* | |
| While an officer interacts with a group of people, they demand to know his badge number. | 2.631 | 0.103 | 0.050 | 0.038** | |
| A person refuses to provide ID to an officer. | 2.323 | -0.004 | 0.048 | 0.939 | |
| A person swears at an officer who initiates a search. | 2.826 | 0.042 | 0.050 | 0.403 | |
| A large crowd of protestors are swearing at an officer and chanting 'defund the police'. | 2.557 | -0.095 | 0.056 | 0.092* | |

**Notes.**This table shows the effect of Sit-D training on officers' tendency to think that subjects are intending to antagonize them (measured in the endline assessment), based on estimating equation (1). Each row is a different regression. One observation is included for each officer. N=1,669. All regressions include stratum fixed effects and additional officer-level covariates (race, gender, experience; as well as baseline values of discretionary arrests, officer injuries and an index of officer activities, key outcomes from the administrative data measured similarly at baseline and endline). The top row shows the results for the Personalization Index, while the remaining rows show the results for the components of the index. Column (1) shows the control mean for each outcome (blank for mean effect indices). Column (2) shows the coefficients on the Sit-D indicator. Column (3) shows robust standard errors. Column (4) shows the observed p-values. Column (5) shows the multiple-inference corrected q-value that adjusts for the false discovery rate across outcomes in a family. The Personalization Index is part of the Navigating Cognitively Demanding Situations Family. *** is significant at the 1% level, ** is significant at the 5% level, and * is significant at the 10% level.

Table B10: Post-FOS Outcomes

| | **Sit-D** (1) | **SE** (2) | **p-value** (3) | **q-value** (4) |
|---|---|---|---|---|
| Recall Index | 0.002 | 0.035 | 0.952 | 0.444 |
| Articulation Index | 0.023 | 0.044 | 0.600 | 0.347 |

**Notes.** This table shows the effect of Sit-D training on officer movement and communication in the FOS exercises (measured in the endline assessment), based on estimating equation (1). Each row is a different regression. One observation is included for each officer. N=1,630. All regressions include stratum fixed effects and additional officer-level covariates (race, gender, experience; as well as baseline values of discretionary arrests, officer injuries and an index of officer activities, key outcomes from the administrative data measured similarly at baseline and endline). Column (1) shows the coefficients on the Sit-D indicator. Column (2) shows robust standard errors. Column (3) shows the observed p-values. Column (4) shows the multiple-inference corrected q-values that adjust for the false discovery rate across outcomes in a family. All outcomes in this table are part of the Officer Performance in the FOS Family. *** is significant at the 1% level, ** is significant at the 5% level, and * is significant at the 10% level.

Table B11: Endline Assessment Outcomes with LASSO-selected Covariates

| | **CM** (1) | **Sit-D** (2) | **SE** (3) | **p-value** (4) | **q-value** (5) | **N** (6) |
|---|---|---|---|---|---|---|
| Knowledge Of Sit-D Concepts Index | - | 0.598 | 0.029 | <0.001*** | 0.001*** | 1,669 |
| Correct assailant level in policy (index) | - | 0.021 | 0.037 | 0.574 | 0.186 | 1,669 |
| Correct force level in policy (index) | - | -0.062 | 0.040 | 0.117 | 0.134 | 1,669 |
| Characterization of assailant who is a direct threat (z-score) | - | 0.078 | 0.048 | 0.105 | 0.134 | 1,669 |
| Total explanations | 3.215 | -0.009 | 0.077 | 0.909 | 0.661 | 1,582 |
| Explanations from multiple categories | 0.667 | 0.042 | 0.023 | 0.072* | 0.094* | 1,582 |
| At least one explanation - assistance category | 0.578 | 0.058 | 0.025 | 0.020** | 0.059* | 1,582 |
| At least one explanation - enforcement category | 0.624 | -0.008 | 0.024 | 0.744 | 0.593 | 1,582 |
| At least one explanation - other category | 0.676 | 0.000 | 0.024 | 0.997 | 0.698 | 1,582 |
| Alternative Features Index (both tasks) | - | 0.099 | 0.032 | 0.002*** | 0.015** | 1,669 |
| Confirming Features Index (both tasks) | - | -0.013 | 0.032 | 0.690 | 0.592 | 1,669 |
| Criminal Interpretations Index (both tasks) | - | -0.053 | 0.025 | 0.035** | 0.072* | 1,669 |
| Decision Time Index (both tasks) | - | -0.062 | 0.032 | 0.052* | 0.074* | 1,669 |
| Processing Time Index (officer-timed task) | - | -0.020 | 0.044 | 0.649 | 0.592 | 1,669 |
| Change - perceived threat & force assessment (index) | - | -0.082 | 0.039 | 0.035** | 0.072* | 1,669 |
| Appropriate actions (index) | - | 0.071 | 0.037 | 0.051* | 0.074* | 1,669 |
| Inappropriate actions (index) | - | -0.005 | 0.033 | 0.875 | 0.661 | 1,669 |
| Personalization Index | - | 0.025 | 0.036 | 0.479 | 0.439 | 1,669 |
| Coping With Stress Index | - | 0.193 | 0.039 | <0.001*** | 0.001*** | 1,669 |
| Emotion Regulation Index | - | 0.078 | 0.029 | 0.007*** | 0.036** | 1,669 |
| Confidence Index | - | 0.099 | 0.041 | 0.016** | 0.059* | 1,669 |
| Did the officer communicate with the person? (index) | - | 0.130 | 0.029 | <0.001*** | 0.001*** | 1,611 |
| Did the officer give verbal direction/ commands to the person? (index) | - | 0.146 | 0.029 | <0.001*** | 0.001*** | 1,611 |
| Did the officer radio dispatch? (index) | - | 0.408 | 0.033 | <0.001*** | 0.001*** | 1,611 |
| Did the officer freeze during the scenario? (index) | - | -0.071 | 0.037 | 0.054* | 0.045** | 1,611 |
| Did the officer kneel or move to cover/ concealment? (index) | - | 0.039 | 0.034 | 0.246 | 0.140 | 1,611 |
| Shooting in the FOS (interaction term) | - | 0.050 | 0.020 | 0.014** | 0.018** | 4,733 |
| Recall Index | - | -0.000 | 0.035 | 0.999 | 0.487 | 1,630 |
| Articulation Index | - | 0.023 | 0.043 | 0.590 | 0.339 | 1,630 |

**Notes.** This table presents estimates of the Sit-D training on key endline assessment outcomes. Each row is a different regression. All regressions include stratum fixed effects and officer-level covariates incorporated by the LASSO double-selection procedure. Column (1) shows the control mean (blank for mean effect indices). Column (2) shows the coefficients on the Sit-D indicator. Column (3) shows robust standard errors. Column (4) shows the observed p-values. Column (5) shows the multiple-inference corrected q-values that adjust for the false discovery rate across outcomes in a family. The top panel shows outcomes in the Knowledge Family, the middle panel shows outcomes in the Navigating Cognitively Demanding Situations Family, and the bottom panel shows outcomes in the Officer Performance in the FOS Family. Column (6) shows the number of observations in each regression. *** is significant at the 1% level, ** is significant at the 5% level, and * is significant at the 10% level.

Table B12: Endline Assessment Outcomes without Additional Covariates

| | **CM** (1) | **Sit-D** (2) | **SE** (3) | **p-value** (4) | **q-value** (5) | **N** (6) |
|---|---|---|---|---|---|---|
| Knowledge Of Sit-D Concepts Index | - | 0.598 | 0.029 | <0.001*** | 0.001*** | 1,669 |
| Correct assailant level in policy (index) | - | 0.021 | 0.037 | 0.574 | 0.219 | 1,669 |
| Correct force level in policy (index) | - | -0.062 | 0.040 | 0.118 | 0.156 | 1,669 |
| Characterization of assailant who is a direct threat (z-score) | - | 0.073 | 0.049 | 0.135 | 0.156 | 1,669 |
| Total explanations | 3.215 | -0.015 | 0.078 | 0.846 | 0.621 | 1,582 |
| Explanations from multiple categories | 0.667 | 0.040 | 0.023 | 0.087* | 0.111 | 1,582 |
| At least one explanation - assistance category | 0.578 | 0.058 | 0.025 | 0.020** | 0.059* | 1,582 |
| At least one explanation - enforcement category | 0.624 | -0.007 | 0.024 | 0.772 | 0.621 | 1,582 |
| At least one explanation - other category | 0.676 | 0.000 | 0.024 | 0.997 | 0.698 | 1,582 |
| Alternative Features Index (both tasks) | - | 0.099 | 0.032 | 0.002*** | 0.016** | 1,669 |
| Confirming Features Index (both tasks) | - | -0.013 | 0.032 | 0.690 | 0.592 | 1,669 |
| Criminal Interpretations Index (both tasks) | - | -0.053 | 0.025 | 0.035** | 0.076* | 1,669 |
| Decision Time Index (both tasks) | - | -0.063 | 0.033 | 0.054* | 0.078* | 1,669 |
| Processing Time Index (officer-timed task) | - | -0.025 | 0.044 | 0.577 | 0.507 | 1,669 |
| Change - perceived threat & force assessment (index) | - | -0.080 | 0.039 | 0.042** | 0.078* | 1,669 |
| Appropriate actions (index) | - | 0.071 | 0.037 | 0.052* | 0.078* | 1,669 |
| Inappropriate actions (index) | - | -0.005 | 0.033 | 0.875 | 0.621 | 1,669 |
| Personalization Index | - | 0.029 | 0.036 | 0.425 | 0.372 | 1,669 |
| Coping With Stress Index | - | 0.189 | 0.039 | <0.001*** | 0.001*** | 1,669 |
| Emotion Regulation Index | - | 0.078 | 0.029 | 0.007*** | 0.036** | 1,669 |
| Confidence Index | - | 0.099 | 0.041 | 0.016** | 0.059* | 1,669 |
| Did the officer communicate with the person? (index) | - | 0.130 | 0.029 | <0.001*** | 0.001*** | 1,611 |
| Did the officer give verbal direction/ commands to the person? (index) | - | 0.146 | 0.029 | <0.001*** | 0.001*** | 1,611 |
| Did the officer radio dispatch? (index) | - | 0.408 | 0.033 | <0.001*** | 0.001*** | 1,611 |
| Did the officer freeze during the scenario? (index) | - | -0.071 | 0.037 | 0.054* | 0.047** | 1,611 |
| Did the officer kneel or move to cover/ concealment? (index) | - | 0.039 | 0.034 | 0.246 | 0.141 | 1,611 |
| Shooting in the FOS (interaction term) | - | 0.050 | 0.022 | 0.022** | 0.029** | 4,733 |
| Recall Index | - | -0.008 | 0.035 | 0.814 | 0.440 | 1,630 |
| Articulation Index | - | 0.020 | 0.044 | 0.642 | 0.380 | 1,630 |

**Notes.** This table presents estimates of the Sit-D training on key endline assessment outcomes. Each row is a different regression. All regressions include stratum fixed effects, but do not include any additional covariates. Column (1) shows the control mean (blank for mean effect indices). Column (2) shows the coefficients on the Sit-D indicator. Column (3) shows robust standard errors. Column (4) shows the observed p-values. Column (5) shows the multiple-inference corrected q-values that adjust for the false discovery rate across outcomes in a family. The top panel shows outcomes in the Knowledge Family, the middle panel shows outcomes in the Navigating Cognitively Demanding Situations Family, and the bottom panel shows outcomes in the Officer Performance in the FOS Family. Column (6) shows the number of observations in each regression. *** is significant at the 1% level, ** is significant at the 5% level, and * is significant at the 10% level.

## B.2 Additional Field Results

In this section, we present additional findings from our field outcomes. All of these tables are referenced in the main text, and some are discussed in greater detail in the sub-sections below.

### Attendance and Attrition

CPD designated Sit-D as a mandatory training, which meant officers assigned to the training were required to complete it. This resulted in relatively high rates of compliance—for example, 990 of 1,059 officers assigned to training completed at least one of the four sessions; 923 completed 2 sessions; and 913 completed both of the first two foundational sessions and at least one of the two applications sessions.

However, compliance was less than 100% for the following reasons. First, at CPD, district commanders can override the order to attend training and cancel trainees out of a class, based on district needs that day. We created make-up classes so officers who missed a session still had other opportunities to complete these classes. Second, officers could be on leave for vacation or medical reasons including injury and illness.[28] In addition, officers could also retire or leave CPD for other reasons. These factors can both reduce training completion and lead to attrition out of the sample as officers who leave no longer appear in the administrative data used for analysis.

Attrition can present a challenge to causal interpretation of the experimental results if it occurs disproportionately in the treatment or control groups. Table B3 examines if treatment assignment predicts attrition. In the top row, we define attrition as occurring if an officer is in our sample but does not appear in any months of administrative data after January 2021, which marks the beginning of the post-training period for most Sit-D officers.

Subsequent rows of Table B3 broaden the definition of attrition. For example, Attrition

[28] An added factor during our training was illness from COVID-19, particularly since much of the training took place prior to the development of the COVID-19 vaccine.

(12 months) equals one if an officer appears in the administrative data in or before February 2021 but no longer appears in the data over March 2021-February 2022, and Attrition (4 months) equals one if the officer does not appear in the last four months of administrative data (November 2021-February 2022). This table shows that treatment does not predict any of these measures, establishing that attrition did not occur disproportionately out of either the training or control groups.

**Effects on Subject Injury and Tactics used by Officers in Use of Force Incidents**

In Table B13 we examine subject injuries and tactics measured from the TRRs. We have limited data to address how Sit-D affects subject injuries since they are not very common and their measurement is noisy (see discussion in Appendix A.2). For example, only 16% of force incidents are reported to result in injury. Table B13 provides some evidence that the training led to reductions in this outcome. The coefficient suggests a 47% fall in officer-reported subject injury. However, this is a large percent reduction from a small base and the effect is not robust to FDR adjustment. In addition, there is no corresponding effect on subject allegations of injury. Table B13 also presents results on hospitalizations (many of which do not arise as a direct consequence of the use of force (see discussion in Appendix A.2); as well as a subset of hospitalizations in which the subject alleged injury or the officer recorded an injury. The coefficients on both of these outcomes are negative, but the effects are imprecise and insignificant at conventional levels. Therefore, this result should be taken as suggestive.

In Table B13, we also examine the types of tactics used by officers in use of force incidents, but do not observe significant effects on this outcome. Given that Sit-D trained officers are involved in fewer use of force incidents, it is possible that they may have used less aggressive tactics (or employed de-escalation tactics) to avoid using force in the first place. But this dynamic is not observable since we only see tactics conditional on an officer being involved in a use of force incident.

### Comparing Main Estimates to DID Estimates

Figure B1 compares our baseline estimates to a difference-in-differences specification, given by:

$$y_{ost} = \alpha_s + \beta SitD_o + \lambda Post_{ot} + \theta(SitD_o \times Post_{ot}) + X_o\delta + \gamma_t + \varepsilon_{ost} \tag{4}$$

where $Post_{ot}$ is the post-training period, and $\theta$ is the difference-in-differences estimate, which captures differential outcomes for Sit-D trained officers in the post-period.

This specification parallels equation (2), which is estimated in Table 3: the post-training period comprises the four months after training completion, while the pre-training period comprises the two years prior to randomization—the period over which baseline controls are measured in equation (2). Covariates again include officer race, gender and experience; and standard errors are clustered on officer.

Estimating equation (4) requires us to have comparable data in outcomes in the baseline and post-training periods. As discussed in the paper (Section 3.2) CPD altered its uses of force classification in March 2020 (utilizing a 4-point use of force classification in our baseline period, and a 3-point classification in our post-training period). The new and old schemes do not line up such that whole categories under the old scheme fall under whole categories of the new scheme.

We create an alignment by hand-categorizing each baseline period use of force incident under the post-period classification. We do so by examining the criteria listed for level 1, 2, and 3 incidents under the March 2020 TRR Directive, and scrutinizing data from different fields in the TRRs (including weapons discharge, hospitalization from injury, etc.) to determine which criteria are met. Since the quality of information across varies across fields, and necessarily involves some subjective assessment, this alignment will not be 100% accurate, and our uses of non-lethal force variable over these two periods will be measured with some noise.

Figure B1 plots the coefficient estimates and 90% CI of $SitD_o \times Post_{ot}$ from equation (4) and $SitD_o$ from equation (2). Qualitatively, the point estimates for the difference-in-differences estimate is larger for discretionary arrests and days off taken for officer injury, but smaller for uses of non-lethal force and the index of officer activities; while the standard errors are larger for all four outcomes. Thus, the difference-in-differences estimates are less precise, which is consistent with low autocorrelation in key outcomes leading this approach to have less power in our setting (McKenzie, 2012). The standard errors may also be large for the uses of non-lethal force, given noise in the alignment process.

Overall, however, the effect sizes in the difference-in-differences specification cannot be said to differ significantly from the effect sizes in our baseline estimates, given the size of the confidence intervals. Based on formal tests of equality from Seemingly Unrelated Estimation, we cannot reject the null hypothesis that the estimates from these two models are the same with p-values of .17, .71, .79 and .23 for uses of non-lethal force, discretionary arrests, officer injuries and officer activities, respectively.

### Allocating Multiple Control Officers to Multiple Post-Treatment Periods

In our primary specification, we randomly allocate control officers to one of seven potential training completion dates (in accordance with our PAP). This ensures that treatment officers and control officers each contribute four post-training months to the data, and that each post-training month contains equal numbers of treated and control observations.

Here we consider an alternate approach which does not rely on random allocation. In this approach, we consider all the different training completion dates (and associated post-training periods) represented among treated officers in a control officer's stratum. We then incorporate the control officer into the post-training dataset for *all* of these "post-training" periods. For example, if there is a stratum where treated officers finished their training on three different dates, then all the control officers in that stratum will contribute three four-monthly periods to the post-training dataset. This approach results in many more

control observations than treatment observations. But, in each period, the number of control observations is proportional to the number of treated observations. Importantly, we continue to cluster standard errors on officer.

Table B18 shows our results under this alternate approach. All the effects remain unchanged, indicating that they do not depend on allocating control officers to particular post-training periods, and are instead robust to utilizing multiple control officers for multiple post-training periods.

**Heterogeneity by Officer Characteristics**

To examine if the training has heterogeneous effects based on characteristics of the officer and districts in which they are employed, we estimate:

$$y_{ost} = \alpha_s + \beta SitD_o + \lambda C_o + \theta(SitD_o \times C_o) + X_o\delta + \gamma_t + \varepsilon_{ost} \tag{5}$$

where $C_o$ is the officer characteristic for which we assess heterogeneity and $\theta$ indicates if there are differential effects based on this characteristic. In Table B19-Table B22 estimates of $\theta$, its standard error, and p-value are reported in columns (4)-(6), respectively.

Since these characteristics may be correlated with other factors that shape adverse policing outcomes, we do not advance a causal interpretation of these analyses, but rather use them to provide suggestive evidence on which types of officers may benefit most from the training.

In the top panel of Table B19, we examine if treatment effects vary based on officer experience, measured as the number of years officers have been on the job. The table shows that reductions in uses of force and discretionary arrests are significantly larger for officers with less experience. The implied differences are substantial. In the control group, there are 55 (50) uses of non-lethal force per month per 1,000 officers among officers with 3 (10) years of experience. The coefficients imply that this outcome falls by 28% among those with 3 years of experience, and 16% among those with 10 years of experience. Similarly,

discretionary arrests are implied to fall by 29% among those with 3 years of experience, but 14% among those with 10 years of experience.

There are two possibilities for why we may observe these effects. Younger officers may learn more from the training, perhaps because they are more malleable and open to new approaches. Or, these officers may face greater need, because they tend to have higher levels of non-lethal force and discretionary arrests.

It does not appear that the extent of learning varies based on experience. For example, Table B20 shows that experienced officers do not have less knowledge of the training or perform worse on any of the assessment outcomes. In fact, these older officers perform *better* on three outcomes, including emotion regulation and identifying disconfirming alternative features of crime scenes. In that regard, the results better align with the idea that Sit-D produces larger effects among less experienced officers because they face a greater need for improvement.

The results in Table B21 are also consistent with this account. Here, we examine heterogeneity based on baseline uses of force and baseline discretionary arrests in the two-year period prior to randomization. The negative coefficients on the interaction terms suggest that the training produces larger effects for those who had higher levels of these adverse policing outcomes prior to training, albeit with varying levels of precision across outcomes.

While these results suggest larger benefits for those with worse outcomes, it is important to note that they do not imply that the benefits are concentrated among a small group of officers. For example, the coefficients in Panel A imply that officers with median values of uses of non-lethal force over the 25 months in the baseline period (i.e., 1.6 force incidents) made 13 percent fewer discretionary arrests in the post-training period (relative to the overall control mean). Similarly, coefficients in Panel B imply that officers making median levels of discretionary arrests in the baseline period (i.e., 2 arrests) engaged in 8 percent fewer uses of force incidents after the training.

We also see similar patterns with those who had a higher rate of making overall arrests,

but do not see differential effects based on the history of complaints over the baseline period. Overall, these heterogeneous effects suggest that officers with a history of using force or making arrests benefit more from Sit-D, perhaps because they face a greater need for training that helps them navigate the situations they are in.

Panels B and C of Table B19 instead examine heterogeneous effects based on officer race and gender. We do not see significant differences based on whether the officer is white, albeit both of the interaction coefficients qualitatively point to larger effects concentrated among officers who are not white. However, non-white officers constitute 54% of the sample (35% are Hispanic, 13% are Black, and the remainder are of other races), so this again does not imply that the training produces effects on a small number of officers. We *do* observe that uses of force decrease significantly more for male officers, as compared to female officers. Male officers constitute 81% of our sample and, in the control mean, have 45 force incidents per 1,000 officers each month—which is three and a half times larger than the rate among female officers. The coefficients in Panel C of Table B19 imply that the training reduces uses of force among male officers by 33%.

Given higher levels of uses of force in the control group among male officers—a point documented by (Ba et al., 2021)—the observed gender heterogeneity is also consistent with Sit-D exerting larger effects among officers with worse starting points, for whom training needs may be greater.

Finally, we consider if the benefits of the training vary based on the level of risk officers face. To do so, in Table B22, we examine heterogeneity based on crime rates in the districts where officers are working. We calculate crimes per 1,000 persons in each district using Chicago Public Data on Crime (provided by CPD) for 2018-2020, scaled by district population data from the Census. Panel A considers the rates of violent crime,[29] and Panel B considers the overall crime rates. Both panels show that there are no significant differential effects based on either measure. This indicates that the effects of Sit-D are similar in

[29]These comprise charges for homicides (FBI Code 01A), criminal sexual assault (02), robbery (03), aggravated assault (04A), and aggravated battery (04B).

different types of districts with different crime rates, and that the benefits of the training are not limited to places where officers face either relatively low levels of risk or relatively high levels of risk.

### Interpreting the Reduction in the Arrests of Black Subjects

In this section, we provide further detailed analyses of Sit-D's effects on racial disparities in arrests and their implications.

First, we consider if the reduction in "other arrests" of Black subjects is driven entirely by other categories of low-level arrests over which officers have discretion—i.e., non-index crime arrests. However, this does not appear to be the case. As we see in Table B25, the effect on non-index crimes is most precisely estimated, but there similar-sized reductions in Black arrests for property and violent crimes, although these latter effects are not statistically significant at conventional level. Thus, it appears that Sit-D leads to broader reductions in racial disparities across multiple levels of arrest categories.

Second, we consider if the reduction in arrests of Black subjects corresponds to a more general reduction in officer activity. Since the vast majority (77%) of arrests in Chicago are of Black subjects, any large reduction in Black arrests will likely correspond to an overall reduction in arrests when pooling across all races. And indeed, this is what we see with "other arrests" in Table B26. However, in this table, none of the other 11 components of officer activity shows a significant decrease, and signs go in different directions. Moreover, the coefficient on the index aggregating these components is positive, albeit insignificant (as discussed in the main paper). Thus it does not appear that the reduction in "other arrests" reflects more general reductions across various categories of officer activities.

The reductions in higher-level arrest categories may raise concerns that the training leads to increases in criminal activity. Prior work suggests that arrests can deter crime (Levitt, 1998), although reductions in lower-level arrests do not lead to crime increases (Cho et al., 2022). Given these findings, we next consider whether Sit-D affects crime rates. Our

experiment is not ideally suited for this purpose since we randomize at the officer level, stratifying by units like district-watches. Thus, the fraction of Sit-D officers in the set of sampled officers does not vary across these units by design. However, we are able to leverage variation in the fraction of Sit-D officers to *total* officers working in district-watches (at the time of randomization). Recall that in a district, the number of *total* officers will differ from the number of *sample* officers since the Sit-D sample comprises only officers who took three prerequisites and have been on the job at least two years.

In Table B27, we regress crimes per 1,000 persons in a given district-watch-month on the Sit-D ratio, while incorporating month fixed effects to account for seasonality in crime. We also cluster the standard errors on district-watch, the level at which the ratio varies.[30]

We estimate effects over a four-month period after most officers had finished their training (January-April 2021) as well as a longer twelve-month post-training period (spanning January-December 2021). We additionally present estimates for four- and twelve-month pre-training periods to examine potential pre-trends.

We find no evidence that a higher ratio of Sit-D officers leads to higher total or violent crime rates in the time period after training completion. In addition, we see no evidence of pre-trends in crime, which is reassuring for this analysis. Since the Sit-D ratio leverages non-experimental variation, these results are necessarily suggestive, but they provide some evidence that the training's reduction in racial disparities of arrests does not coincide with an increase in crime.

Moreover, as prior research shows, arrests should not be seen as a proxy for *productivity* with respect to public safety (Rivera and Ba, 2022; Ba et al., 2022; Lum and Nagin, 2017). Note also that Americans are increasingly likely to say that the criminal justice system is unfair, and many believe that reducing racial bias should be a priority (Brenan, 2023). And officers may have multiple goals that they are weighing, including the harm these arrests might cause and the hassle they entail. Our data therefore do not speak to whether this

[30]Data on crime come from CPD's Citizen Law Enforcement Analysis and Reporting system (CLEAR).

reduction in arrests is desirable or undesirable. But it does appear that Sit-D leads to a broad reduction in racial disparities of discretionary and other arrests.

**Measuring Spillovers**

Randomizing officers into the Sit-D training within unit x watches could lead to potential spillover effects. The presence of such spillovers will lead us to underestimate the treatment effect, if control officers lower the effects on treatment officers, or if treatment officers boost the performance of control officers.

To examine potential spillovers, we leverage the ratio of Sit-D officers to the total number of officers working in a district (at the time of randomization). As discussed in the previous sub-section of the appendix, randomizing within unit-watches means that there is no variation in the ratio of Sit-D officers to officers in the sample within each stratum. However, there is variation in the ratio of Sit-D officers to total officers in the stratum since sample officers are a subset of total officers.

Analytically, we regress the field outcomes on the Sit-D indicator and its interaction with this Sit-D ratio. Specifically, we estimate:

$$y_{ost} = \alpha_s + \beta SitD_o + \theta(SitD_o \times Ratio_s) + X_o\delta + \gamma_t + \varepsilon_{ost} \tag{6}$$

where $y_{ost}$ are outcomes for officer $o$ in stratum $s$ in month $t$; and $Ratio_s$ is the ratio of Sit-D officers in stratum $s$.

This specification mirrors our core analytical approach as reflected in equations (2) and (5). We examine outcomes at the officer level in the focal period, four months after the training, and incorporate time and stratum fixed effects, as well as our core officer-level controls. However, we cluster the standard errors at the more conservative stratum-level, since the Sit-D ratio only varies at this higher level.

Notice that the sign on the Sit-D x Ratio coefficient tell us the direction of the spillover

effects. If untrained officers lower the effectiveness of Sit-D training among treated officers, we should observe a larger share of treated officers reinforcing or leading to larger treatment effects. In other words, the sign of this coefficient should be negative for the adverse policing outcomes. Conversely, if Sit-D officers produce positive spillovers on untrained officers, a higher share of treated officers should be associated with smaller treatment effects—i.e., the sign on this coefficient will then be positive for these outcomes.

Table B28 presents our estimates of equation (6). We observe substantial spillover effects in the two adverse policing outcomes, as reflected in the significant coefficient on the Sit-D x Ratio variable (see main text for a discussion of magnitudes.) Moreover, the sign is negative, indicating that a higher fraction of Sit-D officers in the stratum reinforce the training's effect. This is consistent with untrained officers lowering Sit-D's effectiveness among trained officers.

Table B13: Auxiliary Outcomes in The Field

| | **CM** (1) | **Sit-D** (2) | **SE** (3) | **p-value** (4) | **q-value** (5) |
|---|---|---|---|---|---|
| **Panel A: Specific Use of Force Measures** | | | | | |
| Uses of force (level 1 only) | 22.618 | -4.215 | 3.547 | 0.235 | |
| Uses of force (level 2 only) | 15.502 | -4.658 | 2.607 | 0.074* | |
| Uses of force (levels 1-3) | 38.374 | -7.420 | 4.614 | 0.108 | |
| Uses of force (levels 2-3 only) | 15.756 | -3.206 | 2.737 | 0.242 | |
| **Panel B: Additional TRR Outcomes** | | | | | |
| Subject injuries (officer reported) | 5.337 | -2.505 | 1.415 | 0.077* | 0.625 |
| Subject allegations of injuries | 6.861 | -0.239 | 1.794 | 0.894 | 1.000 |
| Hospitalization | 15.756 | -2.938 | 2.541 | 0.248 | 0.855 |
| Hospitalizations and either subject alleged injury or officer reported an injury | 8.132 | -1.729 | 1.833 | 0.346 | 0.855 |
| Tactics used in uses of force incidents (index) | - | 0.001 | 0.009 | 0.866 | 1.000 |

**Notes.** This table shows the effect of Sit-D training on auxiliary field outcomes based on estimating equation (2). Each row is a separate regression. Four monthly post-training observations are included for each officer. N=8,070. All regressions iinclude stratum fixed effects, month fixed effects, and additional officer-level covariates (race, gender, experience; as well as baseline values of discretionary arrests, officer injuries and an index of officer activities, key outcomes from the administrative data measured similarly at baseline and endline). Panel A presents effects on specific use of force measures from the TRRs, and Panel B presents effects on additional outcomes from the TRRs. Outcomes are measured per 1,000 officers per month, except for tactics used in uses of force, which is measured per officer per month. Column (1) shows the control mean for each outcome (blank for mean effect indices). Column (2) presents the coefficient on the Sit-D indicator from equation (2). Column (3) shows the standard errors, clustered on officer, and column (4) shows the observed p-value. Column (5) presents the multiple-inference corrected q-values that adjust for the false discovery rate across outcomes in a family. Subject injuries, subject allegations of injuries, the hospitalization variables and the tactics variables constitute the Auxiliary TRR family. *** is significant at the 1% level , ** is significant at the 5% level, and * is significant at the 10% level.

Table B14: Downstream Consequences from Officers' Actions

| | **CM** (1) | **Sit-D** (2) | **SE** (3) | **p-value** (4) | **q-value** (5) |
|---|---|---|---|---|---|
| Commendations and awards | 636.595 | -18.739 | 32.490 | 0.564 | 1.000 |
| Total complaints | 35.578 | -2.706 | 4.552 | 0.552 | 0.433 |
| Force and abuse related complaints (index) | - | -0.011 | 0.013 | 0.415 | 1.000 |

**Notes.** This table presents estimates of the Sit-D training on additional outcomes that are downstream from an officers' actions, based on estimating equation (2). Each row is a separate regression. Four monthly post-training observations are included for each officer. N=8,070. All regressions include stratum fixed effects, month fixed effects, and additional officer-level covariates (race, gender, experience; as well as baseline values of discretionary arrests, officer injuries and an index of officer activities, key outcomes from the administrative data measured similarly at baseline and endline). Outcomes are measured per 1,000 officers per month, except for force and abuse related complaints, which is measured per officer per month. Column (1) shows the control mean for each outcome (blank for mean effect indices). Column (2) presents the coefficient on the Sit-D indicator from equation (2). Column (3) shows the standard errors, clustered on officer, and column (4) shows the observed p-value. Column (5) presents the multiple-inference corrected q-values that adjust for the false discovery rate across outcomes in a family. The outcomes in this table constitute the Downstream Actions from Officers' Actions Family. *** is significant at the 1% level , ** is significant at the 5% level, and * is significant at the 10% level.

Table B15: Effects on Arrests for Non-index Crimes

| | **CM** (1) | **Sit-D** (2) | **SE** (3) | **p-value** (4) |
|---|---|---|---|---|
| Non-index arrests | 1945.870 | -171.110 | 108.775 | 0.116 |
| Chicago municipal code violations | 47.776 | -16.761 | 7.513 | 0.026** |
| Criminal sexual abuse | 2.033 | 2.686 | 1.579 | 0.089* |
| Driving under the influence | 42.440 | -2.371 | 7.903 | 0.764 |
| Drug abuse | 547.395 | -74.841 | 58.207 | 0.199 |
| Gambling | 7.878 | -7.684 | 3.658 | 0.036** |
| Liquor license | 0.254 | -0.027 | 0.358 | 0.940 |
| Mob action, loitering and disorderly offenses | 32.529 | -0.319 | 5.276 | 0.952 |
| Offenses against family | 4.066 | -1.226 | 1.466 | 0.403 |
| Prostitution | 2.795 | 0.877 | 1.816 | 0.629 |
| Traffic offenses | 139.009 | -0.910 | 15.237 | 0.952 |
| Warrant | 388.818 | -25.087 | 26.872 | 0.351 |
| Weapon violations | 600.000 | -53.031 | 50.057 | 0.290 |
| Miscellaneous non-index offenses | 130.877 | 7.585 | 10.267 | 0.460 |

**Notes.** This table shows the effect of Sit-D training on arrests for Non-index crimes (as defined by Rivera and Ba (2022)), based on estimating equation (2). Each panel is a separate regression. Four monthly post-training observations are included for each officer. N=8,070. All regressions include stratum fixed effects, month fixed effects, and additional officer-level covariates (race, gender, experience; as well as baseline values of discretionary arrests, officer injuries and an index of officer activities, key outcomes from the administrative data measured similarly at baseline and endline). The top row presents the sum of all arrests classified as non-index crimes, while remaining rows separately show effects individually on each FBI charge category constituting the sum. Outcomes are measured per 1,000 officers per month. Column (1) shows the control mean for each outcome. Column (2) presents the coefficient on the Sit-D indicator from estimating equation (2). Column (3) shows the standard errors, clustered on officer, and column (4) shows the observed p-value. *** is significant at the 1% level , ** is significant at the 5% level, and * is significant at the 10% level.

Table B16: Field Outcomes Three Months after the Training

| | **CM** (1) | **Sit-D** (2) | **SE** (3) | **p-value** (4) | **q-value** (5) |
|---|---|---|---|---|---|
| Uses of non-lethal force | 37.225 | -9.388 | 4.916 | 0.056* | 0.060* |
| Discretionary arrests | 37.563 | -10.869 | 4.978 | 0.029** | 0.060* |
| Officer injuries (days off) | 1.161 | -0.633 | 0.176 | 0.0003*** | 0.001*** |
| Officer activities (index) | - | 0.016 | 0.015 | 0.290 | 0.170 |

**Notes.** This table shows the effect of Sit-D training on key field outcomes in a hypothetical alternative focal period three months after the training. Each row is a separate regression. Three monthly post-training observations are included for each officer. N=6,067. All regressions include stratum fixed effects, month fixed effects, and additional officer-level covariates (race, gender, experience; as well as baseline values of discretionary arrests, officer injuries and an index of officer activities, key outcomes from the administrative data measured similarly at baseline and endline). Outcomes are measured per 1,000 officers per month, except officer injuries, which is measured per officer per month. Column (1) shows the control mean for each outcome (blank for mean effect indices). Column (2) presents the coefficient on the Sit-D indicator from equation (2). Column (3) shows the standard errors, clustered on officer, and column (4) shows the observed p-value. Column (5) presents the multiple-inference corrected q-values that adjust for the false discovery rate across outcomes in a family. Uses of non-lethal force and discretionary arrests constitute the Adverse Policing Outcomes Family, and officer injuries and the officer activities index constitute the Officer Safety and Activity Family. *** is significant at the 1% level , ** is significant at the 5% level, and * is significant at the 10% level.

Table B17: Key Field Outcomes - Robustness to Controls

| | **CM** (1) | **Sit-D** (2) | **SE** (3) | **p-value** (4) | **q-value** (5) |
|---|---|---|---|---|---|
| **Panel A: LASSO-selected Covariates** | | | | | |
| Uses of non-lethal force | 38.119 | -8.688 | 4.608 | 0.059* | 0.064* |
| Discretionary arrests | 36.849 | -8.128 | 4.312 | 0.059* | 0.064* |
| Officer injuries (days off) | 1.179 | -0.590 | 0.177 | 0.0008*** | 0.002*** |
| Officer activities (index) | - | 0.026 | 0.020 | 0.180 | 0.100* |
| **Panel B: No Additional Covariates** | | | | | |
| Uses of non-lethal force | 38.119 | -7.834 | 4.630 | 0.091* | 0.100* |
| Discretionary arrests | 36.849 | -7.784 | 4.391 | 0.076* | 0.100* |
| Officer injuries (days off) | 1.179 | -0.590 | 0.177 | 0.0008*** | 0.002*** |
| Officer activities (index) | - | 0.033 | 0.020 | 0.101 | 0.054* |

**Notes.** This table shows the effect of Sit-D training on key field outcomes, varying the control set. Each row is a separate regression. Four monthly post-training observations are included for each officer. N=8,070. All regressions include stratum fixed effects and month fixed effects. Regressions in Panel A include officer-level covariates incorporated by the LASSO double-selection procedure. Regressions in Panel B do not include any additional covariates. Outcomes are measured per 1,000 officers per month, except officer injuries, which is measured per officer per month. Column (1) shows the control mean for each outcome (blank for mean effect indices). Column (2) presents the coefficient on the Sit-D indicator from estimating equation (2). Column (3) shows the standard errors, clustered on officer, and column (4) shows the observed p-value. Column (5) presents the multiple-inference corrected q-values that adjust for the false discovery rate across outcomes in a family. Uses of non-lethal force and discretionary arrests constitute the Adverse Policing Outcomes Family, and officer injuries and the officer activities index constitute the Officer Safety and Activity Family. *** is significant at the 1% level, ** is significant at the 5% level, and * is significant at the 10% level.

Table B18: Alternate Allocation of Control Officers to Post-training Periods

| | **CM** (1) | **Sit-D** (2) | **SE** (3) | **p-value** (4) | **q-value** (5) |
|---|---|---|---|---|---|
| Uses of non-lethal force | 43.844 | -8.513 | 4.208 | 0.043** | 0.089* |
| Discretionary arrests | 42.965 | -6.777 | 3.885 | 0.081* | 0.089* |
| Officer injuries (days off) | 1.267 | -0.551 | 0.173 | 0.001*** | 0.003*** |
| Officer activities (index) | - | 0.029 | 0.025 | 0.240 | 0.137 |

**Notes.** This table shows the effect of Sit-D training on key field outcomes using a specification in which each control officer is incorporated into the dataset for all the post-training periods represented among all the treated officers in their stratum. Each row is a separate regression. Four monthly post-training observations are included for each treatment officer, but more than four monthly post-training observations are included for each control officer. N=12,095. All regressions include stratum fixed effects, month fixed effects, and additional officer-level covariates (race, gender, experience; as well as baseline values of discretionary arrests, officer injuries and an index of officer activities, key outcomes from the administrative data measured similarly at baseline and endline). Outcomes are measured per 1,000 officers per month, except officer injuries, which is measured per officer per month. Column (1) shows the control mean for each outcome (blank for mean effect indices). Column (2) presents the coefficient on the Sit-D indicator from estimating equation (2). Column (3) shows the standard errors, clustered on officer, and column (4) shows the observed p-value. Column (5) presents the multiple-inference corrected q-values that adjust for the false discovery rate across outcomes in a family. Uses of non-lethal force and discretionary arrests constitute the Adverse Policing Outcomes Family, and officer injuries and the officer activities index constitute the Officer Safety and Activity Family. *** is significant at the 1% level , ** is significant at the 5% level, and * is significant at the 10% level.

Table B19: Effects on Field Outcomes by Officer Experience, Race, and Gender

**Panel A: Effects on Field Outcomes by Officer Experience**

| | Sit-D | | | Sit-D × Experience | | |
|---|---|---|---|---|---|---|
| | Coef | SE | p-value | Coef | SE | p-value |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Uses of non-lethal force | -18.958 | 8.417 | 0.024** | 1.104 | 0.551 | 0.045** |
| Discretionary arrests | -22.852 | 8.030 | 0.004*** | 1.574 | 0.521 | 0.003*** |

**Panel B: Effects on Field Outcomes by Officer Race**

| | Sit-D | | | Sit-D × White | | |
|---|---|---|---|---|---|---|
| | Coef | SE | p-value | Coef | SE | p-value |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Uses of non-lethal force | -14.872 | 6.770 | 0.028** | 13.133 | 9.322 | 0.159 |
| Discretionary arrests | -13.390 | 6.113 | 0.029** | 10.763 | 8.558 | 0.209 |

**Panel C: Effects on Field Outcomes by Officer Gender**

| | Sit-D | | | Sit-D × Male | | |
|---|---|---|---|---|---|---|
| | Coef | SE | p-value | Coef | SE | p-value |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Uses of non-lethal force | 14.497 | 8.443 | 0.086* | -29.056 | 10.117 | 0.004*** |
| Discretionary arrests | 4.904 | 8.673 | 0.572 | -16.633 | 10.213 | 0.104 |

**Notes.** This table presents heterogeneous effects of the Sit-D training by officer experience (Panel A), race (Panel B), and gender (Panel C), based on estimating equation (5). Each row is a separate regression. Four monthly post-training observations are included for each officer. N=8,070. All outcomes are measured per 1,000 officers per month. All regressions include stratum fixed effects, month fixed effects, and additional officer-level covariates (race, gender, experience; as well as baseline values of discretionary arrests, officer injuries and an index of officer activities, key outcomes from the administrative data measured similarly at baseline and endline). Columns (1)-(3) show the coefficient, standard error and p-value for estimates of the Sit-D indicator. Columns (4)-(6) show the coefficient, standard error and p-value for estimates of Sit-D interacted with officer experience in Panel A, race in Panel B, and gender in Panel C. Standard errors are clustered on officer. *** is significant at the 1% level, ** is significant at the 5% level, and * is significant at the 10% level.

Table B20: Effects on Endline Assessment Outcomes by Officer Experience

| | Sit-D | | | Sit-D × Experience | | | |
|---|---|---|---|---|---|---|---|
| | Coef (1) | SE (2) | p-value (3) | Coef (4) | SE (5) | p-value (6) | N (7) |
| Knowledge Of Sit-D Concepts Index | 0.585 | 0.047 | 0.6424 | 0.001 | 0.004 | 0.747 | 1,669 |
| Correct assailant level in policy (index) | -0.084 | 0.060 | 0.166 | 0.012 | 0.005 | 0.016** | 1,669 |
| Correct force level in policy (index) | -0.107 | 0.065 | 0.098* | 0.005 | 0.006 | 0.422 | 1,669 |
| Characterization of assailant who is a direct threat (z-score) | 0.093 | 0.076 | 0.225 | -0.001 | 0.007 | 0.891 | 1,669 |
| Total explanations | -0.021 | 0.128 | 0.869 | 0.001 | 0.011 | 0.937 | 1,582 |
| Explanations from multiple categories | 0.071 | 0.037 | 0.056* | -0.003 | 0.003 | 0.337 | 1,582 |
| At least one explanation - assistance category | 0.075 | 0.040 | 0.063* | -0.002 | 0.004 | 0.594 | 1,582 |
| At least one explanation - enforcement category | 0.007 | 0.038 | 0.844 | -0.002 | 0.003 | 0.623 | 1,582 |
| At least one explanation - other category | -0.008 | 0.039 | 0.842 | 0.001 | 0.003 | 0.794 | 1,582 |
| Alternative Features Index (both tasks) | 0.013 | 0.051 | 0.799 | 0.010 | 0.004 | 0.032** | 1,669 |
| Confirming Features Index (both tasks) | -0.016 | 0.052 | 0.762 | 0.000 | 0.004 | 0.954 | 1,669 |
| Criminal Interpretations Index (both tasks) | -0.088 | 0.042 | 0.037** | 0.004 | 0.004 | 0.280 | 1,669 |
| Decision Time Index (both tasks) | -0.102 | 0.052 | 0.051* | 0.004 | 0.004 | 0.330 | 1,669 |
| Processing Time Index (officer-timed task) | -0.038 | 0.067 | 0.570 | 0.002 | 0.006 | 0.768 | 1,669 |
| Change - perceived threat & force assessment (index) | -0.105 | 0.063 | 0.097* | 0.003 | 0.005 | 0.578 | 1,669 |
| Appropriate actions (index) | 0.026 | 0.060 | 0.665 | 0.005 | 0.005 | 0.357 | 1,669 |
| Inappropriate actions (index) | -0.024 | 0.053 | 0.657 | 0.002 | 0.005 | 0.697 | 1,669 |
| Personalization Index | -0.026 | 0.057 | 0.650 | 0.006 | 0.005 | 0.274 | 1,669 |
| Coping With Stress Index | 0.249 | 0.062 | 0.6424 | -0.005 | 0.005 | 0.324 | 1,669 |
| Emotion Regulation Index | 0.006 | 0.047 | 0.898 | 0.008 | 0.004 | 0.062* | 1,669 |
| Confidence Index | 0.133 | 0.066 | 0.043** | -0.004 | 0.006 | 0.456 | 1,669 |
| Did the officer communicate with the person? (index) | 0.132 | 0.043 | 0.002*** | -0.001 | 0.004 | 0.903 | 1,611 |
| Did the officer give verbal direction/ commands to the person? (index) | 0.153 | 0.043 | 0.6424 | -0.001 | 0.004 | 0.827 | 1,611 |
| Did the officer radio dispatch? (index) | 0.429 | 0.053 | 0.6424 | -0.002 | 0.005 | 0.623 | 1,611 |
| Did the officer freeze during the scenario? (index) | -0.123 | 0.063 | 0.051* | 0.006 | 0.005 | 0.269 | 1,611 |
| Did the officer kneel or move to cover/ concealment? (index) | 0.020 | 0.055 | 0.708 | 0.002 | 0.005 | 0.667 | 1,611 |
| Recall Index | 0.053 | 0.058 | 0.357 | -0.006 | 0.005 | 0.281 | 1,630 |
| Articulation Index | 0.008 | 0.073 | 0.916 | 0.002 | 0.006 | 0.785 | 1,630 |
| | **Sit-D × Direct Threat** | | | **Sit-D × Direct Threat × Experience** | | | |
| Shooting in the FOS (interaction term) | 0.069 | 0.035 | 0.050** | -0.002 | 0.003 | 0.489 | 4,733 |

**Notes.** This table presents heterogeneous effects of the Sit-D training on endline assessment outcomes by officer experience. Each row is a different regression. One observation is included for each officer. All regressions include stratum fixed effects and additional officer-level covariates (race, gender, experience; as well as baseline values of discretionary arrests, officer injuries and an index of officer activities, key outcomes from the administrative data measured similarly at baseline and endline). For all outcomes except Shooting in the FOS, columns (1)-(3) show the coefficient, standard error, and p-value for estimates of the Sit-D indicator. Columns (4)-(6) show the coefficient, standard error, and p-value for estimates of Sit-D interacted with experience. For shooting in the FOS, columns (1)-(3) show the coefficient, standard error, and p-value estimates for Sit-D interacted with whether the subject presents a direct threat, and columns (4)-(6) show the coefficient, standard error, and p-value estimates for Sit-D interacted with whether the subject presents a direct as well as years of officer experience. Standard errors are robust. Column (7) shows the number of observations in each regression. *** is significant at the 1% level, ** is significant at the 5% level, and * is significant at the 10% level.

Table B21: Effects by Baseline Measures of Field Outcomes

**Panel A: Effects on Field Outcomes by Baseline Use of Non-lethal Force**

| | Sit-D | | | Sit-D × Uses of Non-lethal Force | | |
|---|---|---|---|---|---|---|
| | Coef (1) | SE (2) | p-value (3) | Coef (4) | SE (5) | p-value (6) |
| Uses of non-lethal force | 0.049 | 5.099 | 0.992 | -4.579 | 3.545 | 0.197 |
| Discretionary arrests | 0.588 | 4.655 | 0.900 | -5.526 | 2.464 | 0.025** |

**Panel B: Effects on Field Outcomes by Baseline Discretionary Arrests**

| | Sit-D | | | Sit-D × Discretionary Arrests | | |
|---|---|---|---|---|---|---|
| | Coef (1) | SE (2) | p-value (3) | Coef (4) | SE (5) | p-value (6) |
| Uses of non-lethal force | 4.261 | 5.720 | 0.456 | -3.617 | 1.725 | 0.036** |
| Discretionary arrests | -0.643 | 4.891 | 0.895 | -2.156 | 1.393 | 0.122 |

**Panel C: Effects on Field Outcomes by Baseline Arrests**

| | Sit-D | | | Sit-D × Arrests | | |
|---|---|---|---|---|---|---|
| | Coef (1) | SE (2) | p-value (3) | Coef (4) | SE (5) | p-value (6) |
| Uses of non-lethal force | 6.377 | 7.146 | 0.372 | -0.097 | 0.054 | 0.071* |
| Discretionary arrests | -0.750 | 6.833 | 0.913 | -0.049 | 0.049 | 0.316 |

**Panel D: Effects on Field Outcomes by Baseline Complaints**

| | Sit-D | | | Sit-D × Complaints | | |
|---|---|---|---|---|---|---|
| | Coef (1) | SE (2) | p-value (3) | Coef (4) | SE (5) | p-value (6) |
| Uses of non-lethal force | -3.304 | 5.161 | 0.522 | -4.343 | 4.408 | 0.325 |
| Discretionary arrests | -3.512 | 4.969 | 0.480 | -4.597 | 3.478 | 0.186 |

**Notes.** This table presents heterogeneous effects of the Sit-D training by baseline measures of field outcomes, measured as the sum of these outcomes over the baseline period, comprising the 25 months prior to randomization. The baseline outcomes examined include uses of non-lethal force (Panel A), discretionary arrests (Panel B), total arrests (Panel C), and total complaints (Panel D). The specification is based on estimating equation (5). Each row is a separate regression. Four monthly post-training observations are included for each officer. N=8,070. All outcomes are measured per 1,000 officers per month. All regressions include the standard covariate set described in the notes to Table 3. Columns (1)-(3) show the coefficient, standard error and p-value for estimates of the Sit-D indicator. Columns (4)-(6) show the coefficient, standard error and p-value for estimates of Sit-D interacted with the baseline measure examined in each panel. Standard errors are clustered on officer. *** is significant at the 1% level, ** is significant at the 5% level, and * is significant at the 10% level.

Table B22: Effects on Field Outcomes by Crime Rate

**Panel A: Effects on Field Outcomes by Violent Crime Rate**

| | Sit-D | | | Sit-D × Violent Crime Rate | | |
|---|---|---|---|---|---|---|
| | Coef | SE | p-value | Coef | SE | p-value |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Uses of non-lethal force | -7.542 | 6.833 | 0.270 | -1.258 | 5.351 | 0.814 |
| Discretionary arrests | -4.604 | 6.704 | 0.492 | -3.400 | 4.730 | 0.472 |

**Panel B: Effects on Field Outcomes by Crime Rate**

| | Sit-D | | | Sit-D × Crime Rate | | |
|---|---|---|---|---|---|---|
| | Coef | SE | p-value | Coef | SE | p-value |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Uses of non-lethal force | -10.786 | 8.255 | 0.192 | 0.183 | 0.725 | 0.801 |
| Discretionary arrests | -5.445 | 7.932 | 0.492 | -0.279 | 0.641 | 0.663 |

**Notes.** This table presents heterogeneous effects of the Sit-D training by crime rates of the district in which an officer is employed, based on estimating equation (5). Panel A considers violent crime rates and Panel B considers overall crime rates, both calculated over 2018-2020, prior to the start of the training. Each row is a separate regression. Four monthly post-training observations are included for each officer. N=8,070. All outcomes are measured per 1,000 officers per month. All regressions include stratum fixed effects, month fixed effects, and additional officer-level covariates (race, gender, experience; as well as baseline values of discretionary arrests, officer injuries and an index of officer activities, key outcomes from the administrative data measured similarly at baseline and endline). Columns (1)-(3) show the coefficient, standard error and p-value for estimates of the Sit-D indicator. Columns (4)-(6) show the coefficient, standard error and p-value for estimates of Sit-D interacted with the district's violent or overall crime rate. Standard errors are clustered on officer. *** is significant at the 1% level, ** is significant at the 5% level, and * is significant at the 10% level.

Table B23: Arrests by Race of Subject

| | **CM** (1) | **Sit-D** (2) | **SE** (3) | **p-value** (4) |
|---|---|---|---|---|
| Discretionary arrests: Black subjects | 31.258 | -8.844 | 3.966 | 0.026** |
| Discretionary arrests: White subjects | 1.779 | -0.123 | 0.888 | 0.890 |
| Discretionary arrests: Hispanic subjects | 3.812 | 0.493 | 1.340 | 0.713 |
| All arrests: Black subjects | 2016.773 | -219.713 | 103.279 | 0.034** |
| All arrests: White subjects | 150.191 | -11.922 | 11.530 | 0.301 |
| All arrests: Hispanic subjects | 435.832 | -10.152 | 31.552 | 0.748 |
| All arrests: All other race subjects | 19.060 | 2.126 | 4.273 | 0.619 |
| Other arrests: Black subjects | 1985.515 | -210.869 | 102.155 | 0.039** |
| Other arrests: White subjects | 148.412 | -11.800 | 11.441 | 0.302 |
| Other arrests: Hispanic subjects | 432.020 | -10.646 | 31.441 | 0.735 |
| Other arrests: All other race subjects | 19.060 | 2.126 | 4.273 | 0.619 |

**Notes.** This table shows heterogeneous effects of the Sit-D training on arrests, based on subject race. The race categories are: Black, White, Hispanic and All other races (which include Asian/Pacific Islander and Native American). Each row is a separate regression, based on estimating equation (2). Four monthly post-training observations are included for each officer. N=8,070. Outcomes are measured per 1,000 officers per month. "Discretionary arrests" comprise our pre-specified categories; "other arrests" comprise all other arrests that do not fall under the discretionary arrests variable; and "all arrests" comprise the sum of discretionary and other arrests, spanning all arrests made in that month. There were no discretionary arrests in the all other races category in our sample, so the table does not include this regression. All regressions include stratum fixed effects, month fixed effects, and additional officer-level covariates (race, gender, experience; as well as baseline values of discretionary arrests, officer injuries and an index of officer activities, key outcomes from the administrative data measured similarly at baseline and endline). Column (1) shows the control mean for each outcome. Column (2) presents the coefficient on the Sit-D indicator from estimating equation (2). Column (3) shows the standard errors, clustered on officer, and column (4) shows the observed p-value. *** is significant at the 1% level, ** is significant at the 5% level, and * is significant at the 10% level.

Table B24: Arrests of Black Subjects and Other Subjects (Z-Scores)

| | **CM** (1) | **Sit-D** (2) | **SE** (3) | **p-value** (4) |
|---|---|---|---|---|
| Discretionary arrests: Black subjects | - | -0.044 | 0.020 | 0.026** |
| Discretionary arrests: Other subjects | - | 0.005 | 0.022 | 0.821 |
| All arrests: Black subjects | - | -0.059 | 0.028 | 0.034** |
| All arrests: Other subjects | - | -0.015 | 0.028 | 0.606 |
| Other arrests: Black subjects | - | -0.057 | 0.028 | 0.039** |
| Other arrests: Other subjects | - | -0.015 | 0.028 | 0.598 |

**Notes.** This table shows heterogeneous effects of the Sit-D training on arrests, for Black subjects and subjects of all other races. Each row is a separate regression, based on estimating equation (2). Four monthly post-training observations are included for each officer. N=8,070. Outcomes (per officer per month) have been converted to Z-scores. "Discretionary arrests" comprise our pre-specified categories; "other arrests" comprise all other arrests that do not fall under the discretionary arrests variable; and "all arrests" comprise the sum of discretionary and other arrests, spanning all arrests made in that month. All regressions include stratum fixed effects, month fixed effects, and additional officer-level covariates (race, gender, experience; as well as baseline values of discretionary arrests, officer injuries and an index of officer activities, key outcomes from the administrative data measured similarly at baseline and endline). Column (1) shows the control mean for each outcome (blank for mean effect indices). Column (2) presents the coefficient on the Sit-D indicator from estimating equation (2). Column (3) shows the standard errors, clustered on officer, and column (4) shows the observed p-value. *** is significant at the 1% level, ** is significant at the 5% level, and * is significant at the 10% level.

Table B25: Violent, Property and Non-index Crime Arrests of Black and Other Subjects

| | **CM** (1) | **Sit-D** (2) | **SE** (3) | **p-value** (4) |
|---|---|---|---|---|
| Violent arrests: Black subjects | 290.470 | -28.238 | 19.092 | 0.139 |
| Violent arrests: Other subjects | 113.088 | -5.260 | 10.204 | 0.606 |
| Property arrests: Black subjects | 203.812 | -26.222 | 16.572 | 0.114 |
| Property arrests: Other subjects | 68.615 | -8.832 | 7.093 | 0.213 |
| Non-index arrests: Black subjects | 1522.490 | -165.253 | 93.332 | 0.077* |
| Non-index arrests: Other subjects | 423.380 | -5.857 | 32.317 | 0.856 |

**Notes.** This table shows heterogeneous effects of the Sit-D training on violent, property and non-index arrests, for Black subjects and subjects of all other races. Each row is a separate regression, based on estimating equation (2). Four monthly post-training observations are included for each officer. N=8,070. Outcomes are measured per 1,000 officers per month. All regressions include stratum fixed effects, month fixed effects, and additional officer-level covariates (race, gender, experience; as well as baseline values of discretionary arrests, officer injuries and an index of officer activities, key outcomes from the administrative data measured similarly at baseline and endline). Column (1) shows the control mean for each outcome. Column (2) presents the coefficient on the Sit-D indicator from estimating equation (2). Column (3) shows the standard errors, clustered on officer, and column (4) shows the observed p-value. *** is significant at the 1% level, ** is significant at the 5% level, and * is significant at the 10% level.

Table B26: Components of the Officer Activities Index

| | **CM** (1) | **Sit-D** (2) | **SE** (3) | **p-value** (4) |
|---|---|---|---|---|
| Officer activities (index) | - | 0.027 | 0.020 | 0.172 |
| ANOVs | 158.069 | 84.745 | 87.324 | 0.332 |
| Citations - Hazard | 821.347 | -174.339 | 131.328 | 0.184 |
| CTA checks | 245.743 | 15.300 | 93.377 | 0.870 |
| Curfew violations | 0.762 | 0.600 | 0.701 | 0.393 |
| Driver stops | 4539.009 | 323.852 | 343.619 | 0.346 |
| ISRs | 1993.139 | 82.414 | 161.070 | 0.609 |
| Other arrests | 2584.498 | -230.947 | 120.209 | 0.055* |
| Parking citations | 2295.299 | 2011.473 | 1469.631 | 0.171 |
| Recovered guns | 343.583 | -43.139 | 41.773 | 0.302 |
| Recovered vehicles | 6.353 | 2.707 | 1.968 | 0.169 |
| Traffic stops | 146.887 | -15.578 | 25.650 | 0.544 |
| Warrants | 23.634 | -1.200 | 6.599 | 0.856 |

**Notes.** This table shows the effect of Sit-D training on components of the officer activities index based on estimating equation (2). Each panel is a separate regression. Four monthly post-training observations are included for each officer. N=8,070. All regressions include stratum fixed effects, month fixed effects, and additional officer-level covariates (race, gender, experience; as well as baseline values of discretionary arrests, officer injuries and an index of officer activities, key outcomes from the administrative data measured similarly at baseline and endline). The top row presents the officer activities index (measured in standard deviation units). The remaining rows of the table present components of this index, which are measured per 1,000 officers per month. Column (1) shows the control mean for each outcome (blank for mean effect indices). Column (2) presents the coefficient on the Sit-D indicator from equation (2). Column (3) shows the standard errors, clustered on officer, and column (4) shows the observed p-value. *** is significant at the 1% level , ** is significant at the 5% level, and * significant at the 10% level.

Table B27: Effects on Crime Outcomes

| | Coef. (1) | SE (2) | p-value (3) | N (4) |
|---|---|---|---|---|
| **Panel A: Crimes per 1,000 persons** | | | | |
| **Pre-Training Periods** | | | | |
| October 2019 - January 2020 | 0.024 | 0.065 | 0.713 | 260 |
| February 2019 - January 2020 | 0.031 | 0.071 | 0.661 | 780 |
| **Post-Training Periods** | | | | |
| January 2021 - April 2021 | 0.002 | 0.046 | 0.960 | 260 |
| January 2021 - December 2021 | 0.009 | 0.047 | 0.849 | 780 |
| **Panel B: Violent Crimes per 1,000 persons** | | | | |
| **Pre-Training Periods** | | | | |
| October 2019 - January 2020 | 0.002 | 0.009 | 0.787 | 260 |
| February 2019 - January 2020 | 0.003 | 0.009 | 0.765 | 780 |
| **Post-Training Periods** | | | | |
| January 2021 - April 2021 | 0.000 | 0.009 | 0.971 | 260 |
| January 2021 - December 2021 | 0.002 | 0.009 | 0.834 | 780 |

**Notes.** This table examines how the ratio of Sit-D officers to total officers in each district-watch at the time of randomization affects its crime rate over various periods of time. The dependent variable in Panel A is crimes per 1,000 persons and in Panel B is violent crimes per 1,000 persons. In each panel, a four-month pre-training period (from October 2019-January 2020) and a twelve-month pre-training period (from February 2019 to January 2020) are examined. In addition, a four-month post-training period (from January 2021-April 2021) and a twelve-month post-training period (from January 2021 to December 2021) are also examined. Each row is a separate regression. Four (twelve) monthly observations are included for each district-watch in the four-month (twelve-month) period regressions. All regressions include month fixed effects. Column (1) shows the coefficient on the Sit-D Ratio. Column (2) presents the standard error, clustered on district-watch. Column (3) presents the p-value and column (4) presents the number of observations. *** is significant at the 1% level, ** is significant at the 5% level, and * significant at the 10% level.

Table B28: Spillover Effects on Key Field Outcomes

| | Sit-D | | | Sit-D × Ratio | | |
|---|---|---|---|---|---|---|
| | Coef | SE | p-value | Coef | SE | p-value |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Uses of non-lethal force | 8.338 | 9.299 | 0.372 | -1.014 | 0.591 | 0.090* |
| Discretionary arrests | 10.884 | 7.253 | 0.137 | -1.139 | 0.419 | 0.008*** |

**Notes.** This table examines spillover effects of Sit-D in the focal period (1-4 months after the training). It presents heterogeneous effects by the ratio of Sit-D officers to total officers in each unit-watch (stratum) at the time of randomization, on the basis of estimating equation (6). Each row is a separate regression. Four monthly post-training observations are included for each officer. N=8,070. All outcomes are measured per 1,000 officers per month, except officer injuries, which is measured per officer per month. All regressions include stratum fixed effects, month fixed effects, and additional officer-level covariates (race, gender, experience; as well as baseline values of discretionary arrests, officer injuries and an index of officer activities, key outcomes from the administrative data measured similarly at baseline and endline). Columns (1)-(3) show the coefficient, standard error and p-value for estimates of the Sit-D indicator. Columns (4)-(6) show the coefficient, standard error and p-value for estimates of Sit-D interacted with the Sit-D ratio. Standard errors are clustered on the unit-watch (stratum). *** is significant at the 1% level , ** is significant at the 5% level, and * is significant at the 10% level.

Table B29: Spillover Effects on Field Outcomes in Additional Periods

**Panel A: Spillover Effects in Months 5-8**

| | | Sit-D | | | Sit-D × Ratio | | |
|---|---|---|---|---|---|---|---|
| | | Coef (1) | SE (2) | p-value (3) | Coef (4) | SE (5) | p-value (6) |
| Uses of non-lethal force | | 16.061 | 11.059 | 0.150 | -0.941 | 0.581 | 0.109 |
| Discretionary arrests | | 13.430 | 9.383 | 0.156 | -1.339 | 0.602 | 0.028** |

**Panel B: Spillover Effects in Months 9-12**

| | | Sit-D | | | Sit-D × Ratio | | |
|---|---|---|---|---|---|---|---|
| | | Coef (1) | SE (2) | p-value (3) | Coef (4) | SE (5) | p-value (6) |
| Uses of non-lethal force | 37.144 | 8.921 | 10.881 | 0.414 | -0.804 | 0.567 | 0.160 |
| Discretionary arrests | 23.542 | 11.380 | 7.939 | 0.155 | -0.629 | 0.450 | 0.166 |

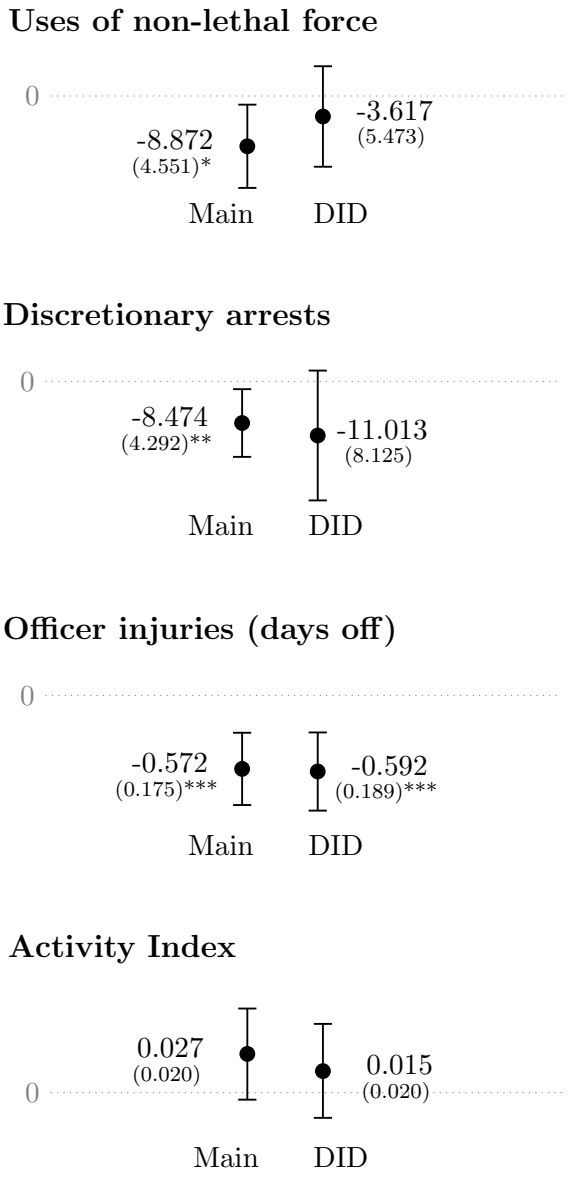
**Notes.** This table examines spillover effects of Sit-D in additional later periods (5-8 months and 9-12 months after the training, in Panels A and B, respectively). It presents heterogeneous effects by the ratio of Sit-D officers to total officers in each unit-watch (stratum) at the time of randomization, on the basis of estimating equation (6). Each row is a separate regression. Four monthly post-training observations are included for each officer, in each panel. N=7,918 in Panel A and N=7,808 in Panel B. All outcomes are measured per 1,000 officers per month, except officer injuries, which is measured per officer per month. All regressions include stratum fixed effects, month fixed effects, and additional officer-level covariates (race, gender, experience; as well as baseline values of discretionary arrests, officer injuries and an index of officer activities, key outcomes from the administrative data measured similarly at baseline and endline). Columns (1)-(3) show the coefficient, standard error and p-value for estimates of the Sit-D indicator. Columns (4)-(6) show the coefficient, standard error and p-value for estimates of Sit-D interacted with the Sit-D ratio. Standard errors are clustered on the unit-watch (stratum). *** is significant at the 1% level , ** is significant at the 5% level, and * is significant at the 10% level.

Table B30: Field Outcomes Twelve Months after the Training

| | **CM** (1) | **Sit-D** (2) | **SE** (3) | **p-value** (4) | **q-value** (5) |
|---|---|---|---|---|---|
| Uses of non-lethal force | 38.151 | -4.528 | 3.161 | 0.152 | 0.087* |
| Discretionary arrests | 32.050 | -5.701 | 2.770 | 0.040** | 0.087* |
| Officer injuries (days off) | 1.290 | -0.121 | 0.171 | 0.479 | 1.000 |
| Officer activities (index) | - | 0.008 | 0.014 | 0.577 | 1.000 |

**Notes.** This table shows the effect of Sit-D training on key field outcomes 12 months after the training. Each row is a separate regression. Twelve monthly post-training observations are included for each officer. N=23,796. All regressions include stratum fixed effects, month fixed effects, and additional officer-level covariates (race, gender, experience; as well as baseline values of discretionary arrests, officer injuries and an index of officer activities, key outcomes from the administrative data measured similarly at baseline and endline). Outcomes are measured per 1,000 officers per month, except officer injuries, which is measured per officer per month. Column (1) shows the control mean for each outcome (blank for mean effect indices). Column (2) presents the coefficient on the Sit-D indicator from equation (2). Column (3) shows the standard errors, clustered on officer, and column (4) shows the observed p-value. Column (5) presents the multiple-inference corrected q-values that adjust for the false discovery rate across outcomes in a family. Uses of non-lethal force and discretionary arrests constitute the Adverse Policing Outcomes Family, and officer injuries and the officer activities index constitute the Officer Safety and Activity Family. *** is significant at the 1% level , ** is significant at the 5% level, and * is significant at the 10% level.

Figure B1: Comparing DID to Main Estimates on Field Outcomes




**Notes.** This figure compares Sit-D's effects on key field outcomes from our main specification (equation (2)) to a DID specification (equation (4)). Each estimate in each panel is a separate regression. All regressions include stratum fixed effects and month fixed effects. Both specifications include additional officer-level covariates (race, gender, experience); while our main specification also includes baseline values of discretionary arrests, officer injuries and an index of officer activities (key outcomes from the administrative data measured in the same way in the baseline and post-training periods). All outcomes are measured per 1,000 officers per month, except officer injuries, which is measured per officer per month. The plot shows the coefficient on SitD from the main specification and the coefficient on SitD x Post from the DID specification, along with 90% Confidence Intervals. Standard errors, clustered on officer, are shown in parentheses. *** is significant at the 1% level, ** is significant at the 5% level, and * is significant at the 10% level. Plots display 90% CI.

# Appendix C: Discussion of Costs and Benefits

Here, we consider (a) whether the costs of Sit-D are in line with other existing police trainings and (b) whether the benefits of Sit-D exceed the costs.

**Costs of training.** We benchmark Sit-D's cost by comparing it to LAPD's Use of Force training, implemented over 2017-2018. The LA training serves as a good comparison to Sit-D since LAPD is another large police department (the third largest in the U.S., following Chicago), and because its Use of Force training uses similar equipment—namely, FOS machines. We focus on its training over 2017-2018, prior to COVID-19, because LAPD suspended its Use of Force training in response to the pandemic. In contrast, note that Sit-D took place during COVID-19, which likely increased implementation cost from a comparative perspective; for example, in the need to schedule more make-up sessions for missed classes owing to higher rates of sick days among police personnel. It is worth noting that this comparison, if anything, will make LAPD's training appear relatively less costly.

As shown in Table C1, we find the cost of the two trainings to be roughly on the same order. We estimate the cost of Sit-D to be $807 per officer assigned to treatment and $864 per officer trained, while the cost of the LAPD's training stands at $715 per officer trained.

The table shows the component costs for teaching personnel and equipment.[31] The bulk of equipment costs stem from the purchase of FOS machines. Upon removing these fixed costs, the recurrent cost of Sit-D stands at $612 -$655 per officer. These recurrent costs are a highly relevant number from the policy perspective of potentially scaling Sit-D, since many police departments already have FOS machines that they use for other types of trainings.

Overall, these estimates suggest that the costs of Sit-D are in line with other relevant police trainings. Importantly, our evaluation demonstrates Sit-D's effectiveness, but we do not yet have direct evidence on the effectiveness of LAPD's training.

Note that annual refresher sessions are added to many police trainings, including CPD's

[31] We do not include officer time in these estimates because officers who were not in Sit-D spent their time in other required trainings. See discussion in Section 3.1.

Use of Force training and the Los Angeles Police Department's De-escalation training. In addition, leading police organizations recommend firearms training three times per year, over four-month intervals (International Association of Law Enforcement Firearms Instructors, 2004, as cited in (Grossi, 2017)). Adding refresher trainings to Sit-D would therefore be aligned with standard approaches to training. Sit-D refresher trainings could, for instance, be one-day sessions that focus on FOS scenarios (similar to the fourth Sit-D session). Based on personnel costs per session, we estimate the cost of this refresher training to be $127-$136 per officer.

**Benefits from reduced officer injuries.** As noted in the main text, there are many non-market benefits of Sit-D, including that the training might increase trust in and cooperation with police departments, and it might reduce the costs that stem from uses of force or low-value arrests. While those benefits are harder to value, we can more readily value the personnel costs saved due to reductions in officer injuries (see the bottom panel of Table C1). Here we find that Sit-D saves $1057 per officer trained in the four months after training alone. Thus, even from this very narrow consideration of potential benefits, we see that the benefits of Sit-D already exceed its costs.

Note that adding a refresher training may increase the benefits of Sit-D by sustaining the effects on outcomes like reduced officer injuries beyond the four-month period. Thus, it would be most appropriate to compare the cost of the core training plus refresher training ($934-$1000 per officer) against the benefit over this sustained period. However, even omitting the potential additional benefit, the cost savings from reduced officer injuries over 4 months ($1057) exceeds the cost of Sit-D inclusive of the refresher training. This underscores the promise of Sit-D as a potential training lever.

Table C1: Cost Analysis

| | LAPD Use of Force | Sit-D |
|---|---|---|
| **Panel A: Costs of Training** | | |
| *Number of Officers:* | | |
| Officers Assigned to Treatment | - | 1,059 |
| Officers Trained | 521 | 990 |
| *Costs:* | | |
| Personnel Cost of Instruction | $244,992 | $536,732 |
| Personnel Cost of Train the Trainer Sessions | $33,151 | $108,310 |
| Equipment Cost | $94,309 | $209,944 |
| *Costs per Officer:* | | |
| Total Cost per Officer Assigned to Treatment | - | $807 |
| Total Cost per Officer Trained | $715 | $864 |
| **Panel B: Benefits from Reduced Officer Injuries** | | |
| Daily Personnel Costs per Officer | - | $462 |
| Four-month Reduction in Days Off Due to Injury | - | 2.288 days |
| Personnel Costs Saved per Officer Trained | - | $1,057 |

**Notes.** Panel A shows the cost estimates for the training. The number of officers trained for LAPD's course reflects the average number of officers trained annually over 2017-2018. The number of officers trained for Sit-D reflects the number of officers who completed at least one of four sessions. The personnel cost of instruction reflects the number of officers and sergeants used to teach the courses, the time they spent teaching the courses, their salary payments, and fringe benefits (estimated to be 38% of annual pay (Bureau of Labor Statistics 2020)). The personnel cost of train-the-trainer sessions reflects the number of hours instructors were trained to be able to teach the courses. Equipment costs for both courses include the cost of Force Option Simulators (FOS), computers, projectors, and basic classroom supplies. LAPD equipment costs also include ammunition for live-fire exercises. Sit-D equipment costs per class are multiplied by three, since three class sessions were run simultaneously. Panel B shows the estimates for one benefit of the training: reduced officer injuries. Personnel costs represent the average daily pay for officers in the training. The four-month reduction in days off due to injury estimate derives from Table 3. Information for these cost estimates come from LAPD, CPD, and public sources. In particular, pay scales for LAPD and CPD are from the City of Los Angeles' MOU No. 24 with the City of Los Angeles Police Protective League, and the City of Chicago's 2020 Classification and Pay Plan, respectively. Pricing information for the FOS equipment reflects actual prices paid by each department, while pricing information for ammunition is from Streichers, a law enforcement and public safety equipment supplier.