The Pearson Institute Discussion Paper No. 2023-13

A Cognitive View of Policing

Oeindrila Dube

Sandy Jo MacArthur

Anuj K. Shah

THE PEARSON INSTITUTE
FOR THE STUDY AND RESOLUTION OF GLOBAL CONFLICTS

A Cognitive View of Policing[*]

Oeindrila Dube[†]
NBER and University of Chicago

Sandy Jo MacArthur[‡]
California Southern University

Anuj K. Shah[§]
University of Chicago

August 30, 2023

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, Ricardo Avalos, Anthony Berglund, Hye Chang, Brian Davis, Hays Golden, 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, UCSD-Rady, West Point, NCGVR, AL CAPONE and the University of Chicago BFI Conference on the Economics of Crime. 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). 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).

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 officers will act without sufficient deliberation and that their actions will be driven by cognitive

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

biases. In this paper, we explore this overlooked perspective and present a cognitive view of policing.

To appreciate this perspective, consider two scenarios. Imagine that an officer encounters a large man in a dark alley who is shouting loudly. The officer sees a glint of something in the man's hand. The officer might conclude that this glint is a weapon, and he might draw (or even fire) his weapon in response. If it turned out that the glint was not a weapon, 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 (i.e., assuming the worst possible outcome), personalizing (i.e., assuming someone is trying to antagonize them), or engaging in confirmation bias (i.e., 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). 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). 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 active duty 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 three courses that are mandatory for all CPD personnel.[2]

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 like excessive force and arbitrary or unnecessary arrests.

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 consider 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

[2] See Section 3.1 for further details.

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. We find that the training leads to reductions in two key adverse outcomes. First, it reduces uses of non-lethal force by 23%.[3] 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.[4] 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. 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. Note that Sit-D produces these effects even though the curriculum does not focus explicitly on racial bias.

These reductions may raise questions about whether the training reduces how active officers are or jeopardizes officer 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 officers take fewer days off from injury on duty in the key four-month period after the training. Thus, Sit-D does not make officers less active and boosts (rather than undermines) officer safety.

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 suggest that the effects diminish over the year, though they do not provide a clear-cut answer to

[3] 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.

[4] 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).

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 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 (see discussion in Section 5), 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 ($1062).

Of course, 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). While some research has shown the benefits of prompting people to be more deliberate, for example by waiting (Imas et al., 2022; Brownback et al., 2023) or slowing down before deciding (Heller et al., 2017), our research goes further by specifying an important element of deliberation: Considering alternative interpretations.

Critically, we situate these broader 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 for example 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.[5] 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 procedural justice during the conversations, and encourage officers to adopt these principles in

[5] 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.

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). These studies highlighted the benefits of slower decision-making, but did not specify exactly how one can become more deliberative. 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. Strikingly, 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. 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.

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 exercises, 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. 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. During the debrief, they 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, officers 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.

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, then officers might be less likely to notice potential risks in the field. This could lead them to use less force, but might also jeopardize officer safety in situations where there are genuine risks. However, we included a mix of scenarios where approximately 60% were threatening and 40% were non-threatening. Thus, it is unlikely that the sample of scenarios alone led officer to see fewer risks in the field. But, the training was designed to get officers to think about many different ways of interpreting the scenarios they encounter, including the possibility that a situation could be less (or more) threatening.

Finally, the Sit-D curriculum does not have any modules specifically on implicit or explicit racial biases in policing. But there are a few places where discussions of racial bias could arise. For instance, the overgeneralization thinking trap warns officers about extrapolating too much from prior situations, including the risks of relying too heavily on stereotypes (encompassing, among others, race-related or age-related stereotypes). Ultimately, however, Sit-D's instruction on racial disparities in policing is fairly limited. Rather, the training might reduce racial disparities if officers are more susceptible to thinking traps when interacting with Black civilians (and thus make greater use of Sit-D content in those situations). Moreover, by making officers more deliberative in general, the training could limit the role that implicit bias plays in shaping officers' actions.

Anonymous Officer Feedback

CPD administered anonymous course evaluations to officers who took Sit-D classes. These evaluations, completed by 942 respondents, indicate that officers generally perceived the training to be highly valuable. For example, when officers were asked "Overall, how much did you like this training?", 83% responded that they either liked the training or liked it a lot. And when they were asked "How useful will this training be to the job you are doing?", 92% responded that it was either very useful or somewhat useful. These responses are encouraging, and suggest that Sit-D classes were generally taught well and that officers found the material to be engaging.

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

[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.

specialized units as the units of assignment.

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 trained officers may influence the outcomes of control officers within a given stratum (and that control officers may influence the outcomes of trained officers). 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. 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] Since Sit-D counted toward fulfilling the 32-hour requirement, for treatment officers it effectively served as a substitute for other existing CPD trainings. Control officers essentially received other CPD trainings in lieu of Sit-D.

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

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

the training by December 2020. Figure 1 presents a consort diagram displaying the timeline.

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]

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).

[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 concepts 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.

[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.

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. 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 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 tasks described above since they were novel for both treatment and control officers. Officers did not debrief these 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 an officer, is injured by an officer, threatens an officer, or physically attacks 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 all non-lethal uses of force, ranging from use of wristlocks to TASER and OC spray (see Appendix A.2 for more detail).[14] We distinguish between lethal and non-lethal levels of force (per our PAP) since there are only 20 lethal force incidents in our sample, and we are not powered to detect changes in this outcome. We therefore focus on non-lethal uses of force, and our main measure is the number of such incidents associated with each officer.

The TRRs contain other information on subject injury and tactics, which we use to construct additional measures analyzed in Table B13. 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.

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

[14] Prior to 2020, the department used a different 4-point grouping which unfortunately does not map cleanly onto the newer 3-point categorization. Since the uses of force measures are not comparable before and after the intervention, we use a measure based on the earlier 4-point grouping only for the purpose of presenting balance statistics.

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 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, the number of these types of discretionary arrests is one of our main measures of adverse policing outcomes.

While we defined one particular subset of arrests that are both discretionary and highly relevant to our training, 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. However, aspects of the training could have implications for racial disparities in policing. For example if officers learn not to overgeneralize, they may be less likely to overgeneralize on the basis of racial stereotypes. Given this implication, we additionally examine if arrest effects differ by race,[15] focusing 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.

[15] We thank a referee for suggesting this.

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 Table 3 alongside our primary administrative outcomes because officer activity and injuries are important for contextualizing the reductions in uses of force and discretionary arrests.

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. Note that there are several issues with our complaints data. As discussed in the appendix, we are not able to identify whether a complaint was generated internally by CPD or by community members, and because of the lag in when complaints are entered into the system, our data are noisy and incomplete. We initially included complaints as part of our primary outcomes in our

[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.

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 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 an 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 .49, 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 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.[17] 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

[17] 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.

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 (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.

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:

$$y_o = \alpha_s + \beta SitD_o + X_o\delta + \varepsilon_o \tag{1}$$

where y_o is the outcome for officer o; 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 α_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.[18] 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 μ_{0k} and σ_{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:

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

where y_{ot} is the outcome for each officer o in each month t; and γ_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 three months, which includes three monthly post-training observations per officer officer; our results are insensitive to this configuration.

[18] 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.

Utilizing CPD's administrative data requires us to demarcate the start of the post-treatment period for officers in both treatment and control. We consider the post-training period to start after officers' last completed class in the Sit-D sequence, since this is when they are meaningfully trained. Since treated officers completed their last class on different dates, their training completion dates—and thus post-training periods—start in different months. Given this data structure, we randomly assign control officers to one of the potential training completion dates, to define their post-treatment periods in an analogous manner. This ensures that we have an even number of treated and control officers in each post-training month of the data. 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 Section 4.2 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:

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

where S_{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 β_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 incorporate additional covariates into the specifications, 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.[19] 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 assess whether Sit-D affected officer outcomes in the field.

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.

[19] Recall that CPD changed how it measured uses of force between baseline and endline which makes them non-comparable across the two periods.

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.[20] 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.[21]

Based on prior work (e.g., Heller et al. (2017)), 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. In fact, we find that Sit-D officers decided on their interpretations significantly

[20] 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.

[21] 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.

faster (fourth row). One possible interpretation of these results is that Sit-D may make officers more efficient both in how they process information and decide on an interpretation. Notice that treatment and control 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) within this same time frame. They appear to have more efficiently processed the scenes.

Perhaps as a result of this efficient processing, Sit-D officers might also be more prepared to subsequently decide on an interpretation. Recall that in this task officers first view the pictures of scenes and then see the two possible interpretations (at which point the timer records how long they take to choose between these possibilities). Sit-D officers may already have considered different potential explanations when viewing the scene—indeed, responses from the Driver's Actions task suggests that this might occur. If trained officers already anticipated these potential interpretations in the first step, they would need to spend less time thinking through them to reach a conclusion in the decision step. In this way, Sit-D officers may also have considered alternatives more efficiently, enabling them to decide faster. 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, when comparing officers' responses to trainers' responses, we find that Sit-D officers listed more appropriate ways of responding to these use of force scenarios (i.e., responses that matched trainers' responses), while 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 a situation 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 Sit-D 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 some responses. However, experimenter demand is unlikely to affect the key measures 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. Measures such these can detect the extent to which officers apply concepts from the course. But since there is no obvious answer that trained officers "should" give, these measures are less susceptible to experimenter demand. As such, the remarkable correspondence in results across a range of measures suggest that the overall results cannot be attributed to experimenter demand alone.

Taken together, these results highlight how Sit-D trains officers to think differently. Most notably, the training improves how officers think through alternative interpretations. Given that this training changes how officers think, 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 focal evaluation period when endline assessments were also conducted). The second column presents the treatment effect, which corresponds to estimates of β 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.[22]

[22] Table B13 also examines other use of force aggregations including lethal force (Level 3) incidents. There are only 20 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.

In the appendix, we examine two other outcomes related to force incidents: subject injuries and tactics. 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.[23] 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.

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

[23] 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) and 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.

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 reduces the aggressiveness of officer responses, and puts them at 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 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.

It also makes clear that the reductions in officer injury cannot reflect lower levels of officer engagement.

Overall, the results in the third and fourth panels of Table 3 suggest that greater deliberation in critical situations, and fewer enforcement actions, do not necessarily endanger officers or stem from officers being less active. If anything, the training appears to alter officer behavior in ways that both keep officers safer and just as active.

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.

Robustness Checks on Field Outcomes. In this section we check the robustness of our main results from Table 3 along four dimensions. We first check the sensitivity of our discretionary arrest results to another classification. We then examine the results in a hypothetical alternative key evaluation period, consider different control sets, and present an alternate specification.

Our discretionary arrests category constitutes a small subset of arrests (covering 2% 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 here for completeness. Non-index arrests are typically for low-level charges of what are called "victimless crimes." However, they are fairly broad in scope, accounting for 69% of all arrests in our sample. Some charges, like municipal code violations, are for crimes like littering and riding a bicycle on the sidewalk. These are most similar in spirit to our discretionary arrests measure, since arrests may not be the most reasonable means of resolving these situations. Other charges, like sexual abuse and weapons violations, are for relatively more serious crimes. Table B15 shows the impact of Sit-D on the sum of all non-index crime arrests in the top row, and on the underlying individual FBI charge categories in the remaining rows.

The table shows that Sit-D leads to a 9% reduction in non-index crime arrests, though the effect is not significant at conventional levels, with a p-value of .108. However, the table also shows that this noisy effect reflects considerable heterogeneity across different charge categories. Sit-D leads to significant increases in arrests for only one non-index category—criminal sexual abuse. In contrast, the training leads to significant reductions in arrests for both gambling and municipal code violations, the latter of which fall by 35% among trained officers. These reductions in the low-level municipal code category align with our account that Sit-D leads officers to reduce unproductive and plausibly unnecessary arrests over which they have discretion.

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. The results show that 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.

Finally, we consider an alternate specification which was not pre-specified but is useful for checking the sensitivity of our results. In our baseline specification, we randomly allocate control officers to one of seven potential training completion dates (per 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 present 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 control officers into the post-training dataset for *all* of these "post-training" periods. Thus, if a control officer is in a stratum where treated officers finished their training on multiple different dates, then that control officer contributes multiple 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.

Effects in Additional 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 months 5-8 and 9-12 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 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, here 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. Thus, while the timing of fade-out for two adverse policing outcomes is unclear given the size of the confidence intervals, the estimates for officer injuries more clearly indicate that fade-out begins to occur five months after the training.

Since treatment effects are strongest in the first four-month interval, and appear to diminish thereafter, Table B19 additionally examines the pooled effect twelve months after the training. We observe a 12% reduction in uses of force and an 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.[24] Thus, though the effects weaken over time, there is still *some* sustained effect on adverse policing outcomes twelve months after the training. Overall, however, the pattern of results shown in Figure 2 and Table B19 suggest that it will be necessary to re-train officers during the year to strongly sustain the training's effect over this duration.

Heterogeneity Analyses. In this section, we consider if the training exerts heterogeneous effects. We first examine if there are differences based on the race of the subject. Since our arrest data contain markers of subject race, in Table 4 we examine discretionary arrests of Black subjects separately from other subjects. The top two rows of the table show that the reduction in this outcome is driven by arrests of Black subjects specifically.

This result suggests that Sit-D may have altered how officers respond to subjects based on their race. It also raises the question: To what extent do these race results generalize across all arrests, rather than discretionary arrests, in particular? To gauge this, the next four rows of Table 4 examine race-disaggregated results for all arrests, and all other arrests besides those we pre-specified as discretionary. For both types of outcomes, there is a substantial and significant reduction in arrests of Black subjects, and a small and insignificant effect on arrests of other subjects.

Three points are worth noting. First, clearly, the base rate of arrests is much higher for Black subjects than other subjects: The control means show that discretionary arrests are 6 times higher while other arrests are 3 times higher for Black subjects than other subjects. However, the effect sizes in percent terms also make clear that the training's effect is larger for those who are Black *even beyond base rate differences.* For example, relative to the control mean, discretionary arrests of Black subjects fall (significantly) by 28%, but (insignificantly) for other subjects by 6%. While the difference is especially large for discretionary arrests, the same race gap is present for other arrests as well. For these other arrests, there is a

[24] 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.

significant 10% reduction for Black subjects (while there is an insignificant 4% reduction for other groups). In addition, formal tests of equality based on Seemingly Unrelated Regressions (SURs) verify that the effects for Black and other subjects are significantly different from one another, for all three sets of arrests outcomes.

In Table B20, we also verify that the null effect on subjects of other races holds when this category is disaggregated into Hispanic, White, and other races.[25] As noted above, Sit-D does not focus extensively on racial bias. However, these findings suggest that teaching officers about broader cognitive biases, and teaching them to be more deliberative in general, could reduce the impact that implicit biases have on their actions, and ultimately, alter racial patterns in policing.

Sit-D's large reduction in "other arrests" of Black subjects, combined with the fact that most such arrests of Chicagoans are of Black Chicagoans (77%), means that we also observe significant decreases in arrests of this category when pooling together all subjects by race (see Table B21). However, as this appendix table also demonstrates, there is no significant reduction in any other component of the officer activities index (or the aggregate index itself), which reinforces the idea that the training does not lead to general reductions in total officer activity.

In a second set of heterogeneity results, in Table B22-Table B24, we examine differences based on the characteristics of officers, and the districts in which they are employed.

We see clear patterns of heterogeneity based on officer experience: Table B22 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 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 the Sit-D has greater impact

[25] 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.

among officers who face worse starting points, for whom training needs may be greater. Appendix B.3 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 B24). 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.

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 Situational Decision-making processed information more efficiently, developed more varied explanations of subject behavior, and updated their assessments to a greater degree as situations became less threatening. 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 the persistent racial disparities in policing (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 $1062 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). Because the most stressful policing situations often involve multiple officers responding at once, peer effects might be particularly influential in determining the extent to which officers deliberate over alternatives before taking action in these moments. On the one hand, this implies that our results may understate the true effect of the training. On the other hand, it may also mean that training a larger fraction of officers will 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., 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.

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.

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.

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.

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.

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

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.

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.

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	**Sit-D**	**SE**	**p-value**	**q-value**
	(1)	(2)	(3)	(4)	(5)
Panel A: Alternative Interpretations of a Subject's Actions					
Total explanations	3.215	-0.014	0.077	0.854	0.597
Explanations from multiple categories	0.667	0.041	0.023	0.075*	0.100*
At least one explanation - assistance category	0.578	0.058	0.025	0.019**	0.064*
At least one explanation - enforcement category	0.624	-0.008	0.024	0.751	0.597
At least one explanation - other category	0.676	-0.000	0.024	0.999	0.699
Panel B: Processing Information and Forming Interpretations					
Alternative Features Index (both tasks)	-	0.102	0.032	0.001***	0.012**
Confirming Features Index (both tasks)	-	-0.014	0.032	0.676	0.573
Criminal Interpretations Index (both tasks)	-	-0.052	0.025	0.040**	0.082*
Decision Time Index (both tasks)	-	-0.062	0.032	0.051*	0.082*
Processing Time Index (officer-timed task)	-	-0.022	0.044	0.611	0.554
Panel C: Use of Force in Dynamic Situations					
Change – perceived threat & force assessment (index)	-	-0.077	0.039	0.047**	0.082*
Appropriate actions (index)	-	0.070	0.037	0.057*	0.082*
Inappropriate actions (index)	-	-0.007	0.033	0.829	0.597

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 officers processed information and formed interpretations in the Pictures Task, and the bottom panel shows how officers responded to use of force scenarios. Each row is a different regression. One observation is included for each officer (N=1,582 for the top panel; N=1,669 for the middle and bottom panels). 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 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-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	0.059*	0.050**
Did the officer kneel or move to cover/ concealment? (index)	0.040	0.033	0.225	0.128

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***	0.051	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	-8.891	4.552	0.051*	0.054*
Discretionary arrests	36.849	-8.477	4.291	0.048**	0.054*
Officer injuries (days off)	1.179	-0.572	0.175	0.001***	0.003***
Officer activities (index)	-	0.021	0.019	0.270	0.157

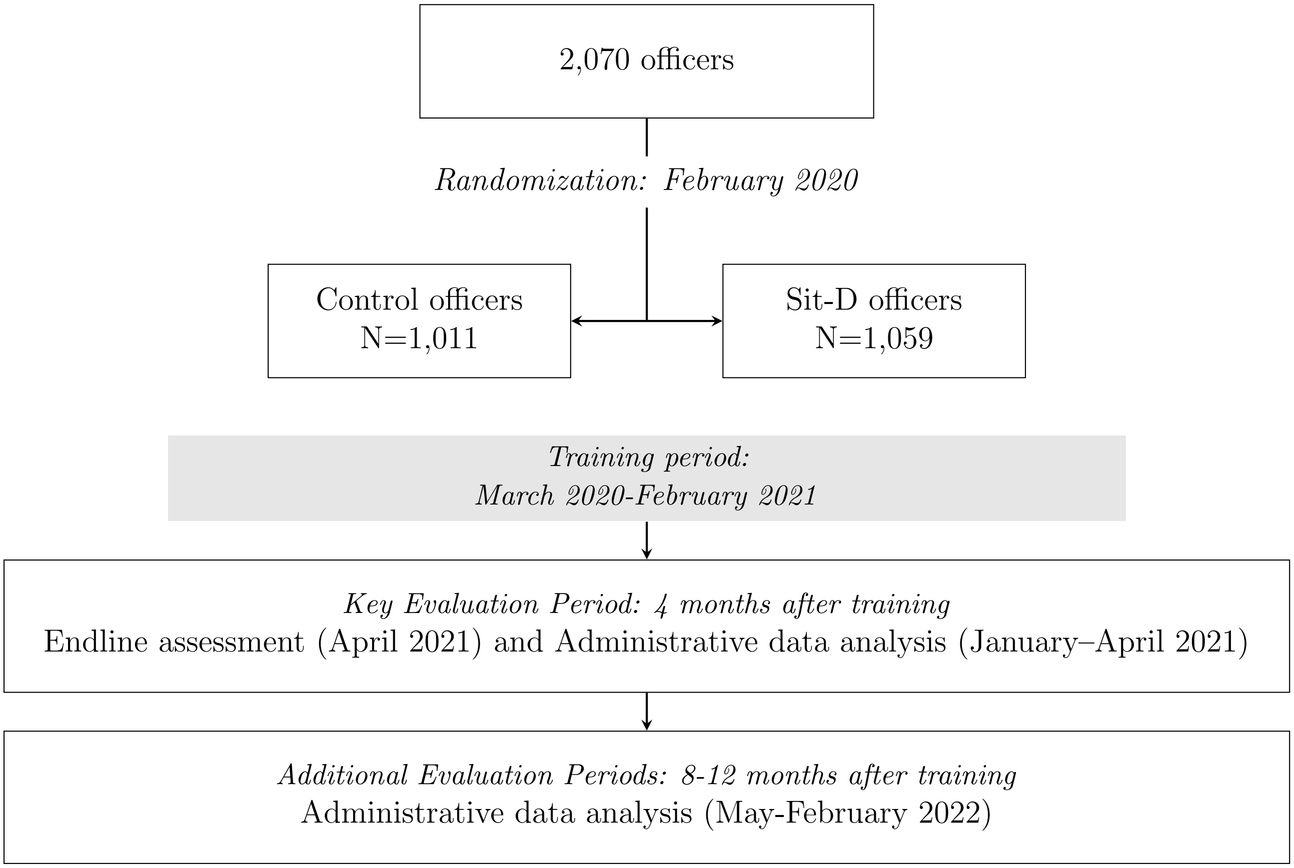
Notes. This table shows the effect of Sit-D training on key field outcomes based on estimating equation (2). Each row is a separate regression. Four monthly post-training observations are included for each officer. N=8,070. All regressions include stratum fixed effects, month fixed effects, and additional 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. 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	-8.821	3.963	0.026**
Discretionary arrests: Other subjects	5.591	0.344	1.637	0.834
All arrests: Black subjects	2016.773	-222.158	103.080	0.031**
All arrests: Other subjects	605.083	-20.727	38.536	0.591
Other arrests: Black subjects	1985.515	-213.338	101.952	0.037**
Other arrests: Other subjects	599.492	-21.071	38.394	0.583

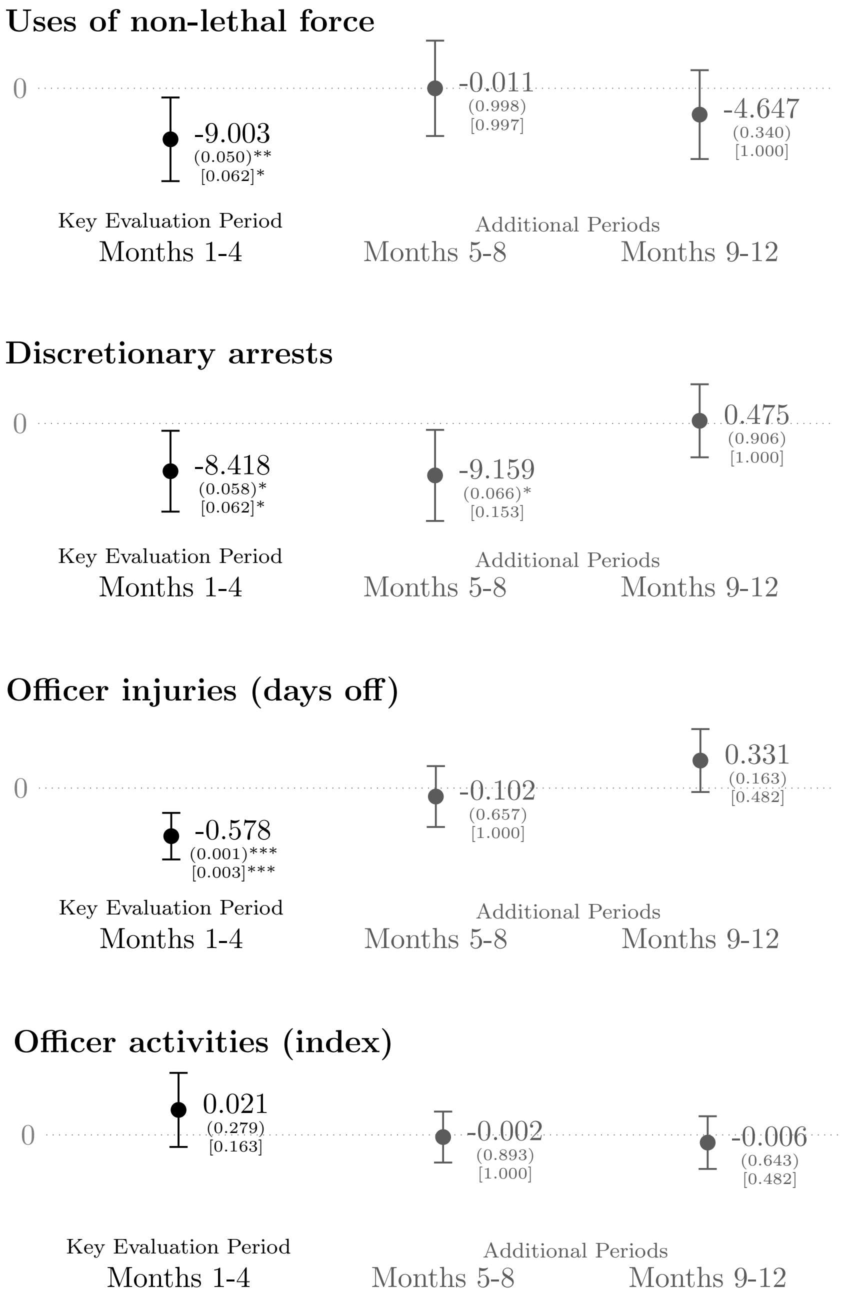
Notes. This table shows heterogenous effects of the Sit-D training on arrests, for black subjects and subjects of all other races (denoted Other Subjects). Each row is a separate regression, as given by equation (2). Four monthly post-training observations are included for each officer. N=8,070. Outcomes 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


2,070 officers
Randomization: February 2020
Control officers
N=1,011
Sit-D officers
N=1,059
Training period:
March 2020-February 2021
Key Evaluation Period: 4 months after training
Endline assessment (April 2021) and Administrative data analysis (January–April 2021)
Additional Evaluation Periods: 8-12 months after training
Administrative data analysis (May-February 2022)

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


Uses of non-lethal force

Discretionary arrests

Officer injuries (days off)

Officer activities (index)

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. One observation is included for each officer-month over twelve months after the training. N=23,796. All regressions include stratum fixed effects and 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 collection 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. 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). Another 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). Another 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; 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. 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 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. 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. Another clip depicted a group of people demanding to know an officer's badge number. Another clip depicted a subject refusing to show their ID to an officer. Another clip depicted a person swearing at an officer as the officer initiates a search. Another 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 intrusion. 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).[26]

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 re-

[26] 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).

spond (e.g., the color of the subject's shoes). And both treatment and control officers wrote 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

Additional 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 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 firerams discharge, use of firearms to deter an animal, and long-range acoustic devices such as sound canons. Level 3 incidents include shootings, as well as using a choke-hold, or using 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 analysis 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 preexisting 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).

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. **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. **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. **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. **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. **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. **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.

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

Table B1: Balance on Key Covariates (Full Sample)

	Control Mean	**Treatment Mean**	**Difference**	**N**
	Panel A: Officer characteristics			
Age	37.840 (8.512)	38.052 (8.653)	0.190 (0.333)	2,070
Years of experience	9.229 (7.397)	9.219 (7.208)	-0.019 (0.277)	2,070
Gender: Male	0.799 (0.401)	0.813 (0.390)	0.015 (0.017)	2,070
Race and ethnicity: Black	0.132 (0.338)	0.125 (0.330)	-0.010 (0.014)	2,070
Race and ethnicity: Hispanic	0.363 (0.481)	0.342 (0.475)	-0.019 (0.021)	2,070
Race and ethnicity: White	0.449 (0.498)	0.476 (0.500)	0.028 (0.021)	2,070
Race and ethnicity: Other	0.056 (0.231)	0.058 (0.233)	0.001 (0.010)	2,070
	Panel B: Officer performance prior to treatment			
Uses of force	1.695 (2.535)	1.600 (2.354)	-0.090 (0.098)	2,070
Uses of force - All but lethal	1.672 (2.499)	1.568 (2.309)	-0.098 (0.096)	2,070
Subject injuries (officer reported)	0.213 (0.626)	0.216 (0.586)	0.003 (0.026)	2,070
Subject allegation of injuries	0.179 (0.514)	0.185 (0.518)	0.006 (0.022)	2,070
Officer injuries (days off)	14.528 (44.546)	14.418 (46.310)	-0.178 (1.980)	2,070
Tactics used in TRRs (index)	0.000 (0.085)	-0.004 (0.071)	-0.004 (0.003)	2,070
Discretionary arrests	3.581 (4.299)	3.665 (4.489)	0.105 (0.152)	2,070
Total officer activities (index)	0.000 (0.232)	0.006 (0.260)	0.008 (0.010)	2,070
Complaints	1.180 (1.705)	1.091 (1.621)	-0.085 (0.068)	2,070
Hospitalizations	0.555 (0.964)	0.520 (0.987)	-0.034 (0.041)	2,070
Awards and commendations	16.616 (16.657)	17.008 (17.709)	0.517 (0.588)	2,070
F-test: p-value = 0.4895				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 (8.165)	38.011 (8.560)	0.432 (0.364)	1,696
Years of experience	9.173 (7.225)	9.204 (7.114)	0.186 (0.306)	1,696
Gender: Male	0.795 (0.404)	0.820 (0.384)	0.024 (0.019)	1,696
Race and ethnicity: Black	0.126 (0.332)	0.126 (0.332)	-0.001 (0.015)	1,696
Race and ethnicity: Hispanic	0.370 (0.483)	0.343 (0.475)	-0.024 (0.023)	1,696
Race and ethnicity: White	0.448 (0.498)	0.472 (0.500)	0.024 (0.024)	1,696
Race and ethnicity: Other	0.056 (0.230)	0.058 (0.235)	0.001 (0.011)	1,696
Panel B: Officer performance prior to treatment				
Uses of force	1.596 (2.311)	1.578 (2.327)	-0.054 (0.103)	1,696
Uses of force - All but lethal	1.573 (2.276)	1.546 (2.274)	-0.061 (0.101)	1,696
Subject injuries (officer reported)	0.199 (0.577)	0.206 (0.553)	0.002 (0.027)	1,696
Subject allegation of injuries	0.162 (0.488)	0.179 (0.509)	0.011 (0.023)	1,696
Officer injuries (days off)	13.167 (42.983)	13.001 (44.358)	-0.671 (2.104)	1,696
Tactics used in TRRs (index)	0.000 (0.082)	-0.003 (0.073)	-0.003 (0.004)	1,696
Discretionary arrests	3.475 (4.198)	3.743 (4.615)	0.183 (0.170)	1,696
Total officer activities (index)	0.000 (0.225)	0.013 (0.285)	0.014 (0.012)	1,696
Complaints	1.077 (1.529)	1.088 (1.669)	0.002 (0.073)	1,696
Hospitalizations	0.529 (0.910)	0.506 (0.960)	-0.031 (0.043)	1,696
Awards and commendations	16.165 (14.563)	17.375 (18.075)	0.933 (0.598)	1,696
F-test: p-value = 0.6419				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 (0.013)	2,070
Attrition (12 months)	0.010	0.011	0.001 (0.004)	2,040
Attrition (8 months)	0.018	0.026	0.008 (0.006)	2,044
Attrition (4 months)	0.045	0.043	-0.002 (0.009)	2,052
Attrition (Endline survey)	0.169	0.192	0.022 (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 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 for 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 above—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)	0.028	0.037	0.451	0.156
Correct force level in policy (index)	-0.065	0.040	0.101	0.113
Characterization of assailant who is a direct threat (z-score)	0.084	0.048	0.080^{*}	0.113

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: Components of Coping with Stress and Emotion Regulation Indices

	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	0.270	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	-	0.078	0.029	0.007***	0.037**
I could be experiencing some emotion and not be conscious of it until some time later.	4.458	0.023	0.057	0.690	
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	0.115	0.070	0.102	

Notes. This table shows the effect of Sit-D training on how officers cope with stress and regulate 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^{**}	0.064^{*}
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	0.056^{*}	
How confident are you in your ability to effectively respond to a shots fired call?	3.614	0.039	0.026	0.128	
How confident are you in your ability to do your job effectively during a protest about policing?	3.634	0.043	0.026	0.095^{*}	
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	0.236	

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	0.442	0.392
While an officer performs a traffic stop, a bystander starts filming.	2.830	0.077	0.047	0.098*	
While an officer interacts with a group of people, they demand to know his badge number.	2.631	0.104	0.050	0.037**	
A person refuses to provide ID to an officer.	2.323	-0.003	0.048	0.947	
A person swears at an officer who initiates a search.	2.826	0.043	0.050	0.395	
A large crowd of protestors are swearing at an officer and chanting "defund the police."	2.557	-0.095	0.056	0.093*	

Notes. This table shows the effect of Sit-D training on officers' tendency to think that subjects intend 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	**SE**	**p-value**	**q-value**
	(1)	(2)	(3)	(4)
Recall Index	0.002	0.035	0.961	0.430
Articulation Index	0.023	0.044	0.601	0.347

Notes. This table shows the effect of Sit-D training on recall of details and articulation of actions following 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 Results on Field Outcomes

In this section, we include tables for auxiliary outcomes in the field, downstream consequences from officers' actions, and robustness checks for the field outcomes. Each of these are discussed in more detail in the main text (see Outcomes in the Field).

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	-4.209	3.551	0.236	
Uses of force (level 2 only)	15.502	-4.682	2.605	0.072*	
Uses of force (levels 1-3)	38.374	-7.430	4.616	0.108	
Uses of force (levels 2-3 only)	15.756	-3.221	2.735	0.239	
Panel B: Additional TRR Outcomes					
Subject injuries (officer reported)	5.337	-2.513	1.413	0.075*	0.607
Subject allegations of injuries	6.861	-0.251	1.793	0.889	1.000
Hospitalization	15.756	-2.938	2.542	0.248	0.837
Hospitalizations and either subject alleged injury or officer reported an injury	8.132	-1.743	1.832	0.342	0.837
Tactics used in uses of force incidents (index)	-	-0.003	0.008	0.692	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. 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	0.637	-0.019	0.032	0.563	1.000
Total complaints	0.036	-0.003	0.005	0.552	1.000
Force and abuse related complaints (index)	-	-0.007	0.013	0.572	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). All outcomes are measured per officer per month. Column (1) shows the control mean for each outcome. 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	-174.224	108.460	0.108
Gambling	7.878	-7.715	3.673	0.036**
Liquor license	0.254	-0.023	0.356	0.949
Prostitution	2.795	0.859	1.814	0.636
Drug abuse	547.395	-76.100	58.148	0.191
Driving under the influence	42.440	-2.655	7.860	0.736
Weapon violations	600.000	-53.162	50.019	0.288
Offenses against family	4.066	-1.215	1.461	0.406
Mob action, loitering and disorderly offenses	32.529	-0.308	5.264	0.953
Chicago municipal code violations	47.776	-16.937	7.527	0.025**
Traffic offenses	139.009	-1.236	15.185	0.935
Warrant	388.818	-25.792	26.812	0.336
Criminal sexual abuse	2.033	2.703	1.591	0.090*
Miscellaneous non-index offenses	130.877	7.358	10.255	0.473

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. 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	-9.395	4.917	0.056*	0.060*
Discretionary arrests	37.563	-10.858	4.977	0.029**	0.060*
Officer injuries (days off)	1.161	-0.632	0.176	0.0003***	0.001***
Officer activities (index)	-	0.012	0.014	0.368	0.226

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. 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	-8.707	4.606	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)	-	0.021	0.019	0.281	0.164
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)	-	0.028	0.020	0.166	0.091*

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. 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	-8.537	4.207	0.043**	0.087*
Discretionary arrests	42.965	-6.814	3.883	0.079*	0.087*
Officer injuries (days off)	1.267	-0.550	0.173	0.001***	0.003***
Officer activities (index)	-	0.029	0.025	0.245	0.140

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. 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: Field Outcomes Twelve Months after the Training

	CM	**Sit-D**	**SE**	**p-value**	**q-value**
	(1)	(2)	(3)	(4)	(5)
Uses of non-lethal force	38.151	-4.583	3.159	0.147	0.082*
Discretionary arrests	32.050	-5.748	2.765	0.038**	0.082*
Officer injuries (days off)	1.290	-0.121	0.171	0.478	1.000
Officer activities (index)	-	0.003	0.012	0.820	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. 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 B20: Arrests by Race of Subject

	CM (1)	**Sit-D** (2)	**SE** (3)	**p-value** (4)
Discretionary arrests: Black subjects	31.258	-8.821	3.963	0.026**
Discretionary arrests: White subjects	1.779	-0.123	0.888	0.890
Discretionary arrests: Hispanic subjects	3.812	0.467	1.342	0.728
All arrests: Black subjects	2016.773	-222.158	103.080	0.031**
All arrests: White subjects	150.191	-12.013	11.527	0.297
All arrests: Hispanic subjects	435.832	-10.754	31.488	0.733
All arrests: All other race subjects	19.060	2.040	4.281	0.634
Other arrests: Black subjects	1985.515	-213.338	101.952	0.037**
Other arrests: White subjects	148.412	-11.891	11.437	0.299
Other arrests: Hispanic subjects	432.020	-11.221	31.380	0.721
Other arrests: All other race subjects	19.060	2.040	4.281	0.634

Notes. This table shows heterogenous 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, as given by 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 B21: Components of the Officer Activities Index

	CM (1)	**Sit-D** (2)	**SE** (3)	**p-value** (4)
Officer activities (index)	-	0.021	0.019	0.270
Other arrests	2584.498	-234.165	119.868	0.051*
PRS, Recovered guns	343.583	-42.962	41.726	0.303
PRS, Recovered vehicles	6.353	2.708	1.967	0.169
PRS, Warrants	23.634	-1.150	6.583	0.861
PRS, Traffic stops	146.887	-16.520	25.330	0.514
PRS, Driver stops	4539.009	311.128	341.509	0.362
PRS, ISRs	1993.139	77.640	160.202	0.628
PRS, ANOVs	158.069	87.559	89.210	0.326
PRS, Citations - Hazard	821.347	-180.566	130.713	0.167
PRS, Curfew violations	0.762	0.598	0.698	0.392
PRS, CTA checks	245.743	15.240	93.414	0.870
PRS, Parking citations	2295.299	2025.611	1469.298	0.168

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.

B.3 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:

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

where C_o is the officer characteristic for which we assess heterogenetiy and θ indicates if there are differential effects based on this characteristic. In Table B22-Table B24 estimates of θ, 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 B22, 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 60 (50) uses of non-lethal force a month per 1,000 officers among officers with 2 (10) years of experience. The coefficients imply that this outcome falls by 28% among those who have been on the force for 2 years, and 16% among those who have been on the job 10 years. Similarly, discretionary arrests are implied to fall by 38% among those with two 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 to be the case that the extent of learning varies based on experience. For example, Table B23 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. This is also consistent with their becoming more active in the field in response to the training (see bottom row of Table B22-Panel A). In that regard, the results are more consistent with the idea that Sit-D produces larger effects among less experienced officers because they face a greater need for improvement.

Panel B of Table B22 instead examines heterogeneous effects based on officer race and gender. We do not see significant differences based on whether the officer is White, but 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 B22 imply that the training reduces uses of force among male officers by 33%.

Given higher levels uses of force in the control group among among male officers—a point documented by (Ba et al., 2021)—the observed gender heterogeniety 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 are localized to places where officers face relatively little risk. To do so, in Table B24, 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 the period 2018-2020, scaled by district population data from the Census. Panel A considers the rates of violent crime,[27] 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 different types of districts with different crime rates, and that the benefits of the training are not limited to places where officers face low levels of risk.

[27] These comprise charges for homicides, manslaughter, assault, robbery, and battery.

Table B22: 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	-18.938	8.411	0.024**	1.100	0.550	0.046**
Discretionary arrests	-22.831	8.030	0.005***	1.572	0.521	0.003***
Officer injuries (days off)	-0.815	0.287	0.005***	0.027	0.024	0.259
Officer activities (index)	-0.036	0.024	0.130	0.006	0.003	0.020**

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	-15.010	6.775	0.027**	13.392	9.329	0.151
Discretionary arrests	-13.457	6.111	0.028**	10.901	8.560	0.203
Officer injuries (days off)	-0.472	0.230	0.040**	-0.218	0.344	0.527
Officer activities (index)	0.036	0.026	0.174	-0.032	0.030	0.290

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	14.627	8.444	0.083*	-29.243	10.119	0.004***
Discretionary arrests	4.982	8.671	0.566	-16.735	10.215	0.102
Officer injuries (days off)	-0.645	0.517	0.213	0.091	0.556	0.870
Officer activities (index)	0.008	0.027	0.762	0.016	0.038	0.673

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 (4). 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 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 B23: 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)	-0.083	0.060	0.169	0.012	0.005	0.016**	1,669
Correct force level in policy (index)	-0.107	0.065	0.100*	0.005	0.006	0.424	1,669
Characterization of assailant who is a direct threat (z-score)	0.093	0.076	0.223	-0.001	0.007	0.892	1,669
Total explanations	-0.022	0.128	0.866	0.001	0.011	0.939	1,582
Explanations from multiple categories	0.071	0.037	0.056*	-0.003	0.003	0.336	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	0.008	0.038	0.838	-0.002	0.003	0.623	1,582
At least one explanation - other category	-0.008	0.039	0.836	0.001	0.003	0.796	1,582
Alternative Features Index (both tasks)	0.013	0.051	0.791	0.010	0.004	0.032**	1,669
Confirming Features Index (both tasks)	-0.016	0.052	0.766	0.000	0.004	0.960	1,669
Criminal Interpretations Index (both tasks)	-0.088	0.042	0.038**	0.004	0.004	0.281	1,669
Decision Time Index (both tasks)	-0.102	0.052	0.050**	0.004	0.004	0.331	1,669
Processing Time Index (officer-timed task)	-0.039	0.067	0.563	0.002	0.006	0.772	1,669
Change - perceived threat & force assessment (index)	-0.105	0.063	0.095*	0.003	0.005	0.578	1,669
Appropriate actions (index)	0.026	0.060	0.661	0.005	0.005	0.360	1,669
Inappropriate actions (index)	-0.023	0.053	0.661	0.002	0.005	0.700	1,669
Personalization Index	-0.025	0.057	0.655	0.006	0.005	0.274	1,669
Coping With Stress Index	0.248	0.062	0.6424	-0.005	0.005	0.327	1,669
Emotion Regulation Index	0.007	0.047	0.887	0.008	0.004	0.063*	1,669
Confidence Index	0.133	0.066	0.044**	-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	0.824	1,611
Did the officer radio dispatch? (index)	0.429	0.053	0.6424	-0.002	0.005	0.624	1,611
Did the officer freeze during the scenario? (index)	-0.123	0.063	0.051*	0.006	0.005	0.267	1,611
Did the officer kneel or move to cover/ concealment? (index)	0.021	0.055	0.705	0.002	0.005	0.664	1,611
Recall Index	0.053	0.058	0.361	-0.006	0.005	0.280	1,630
Articulation Index	0.008	0.073	0.915	0.002	0.006	0.786	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	0.490	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 threat 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 B24: 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 (1)	SE (2)	p-value (3)	Coef (4)	SE (5)	p-value (6)
Uses of non-lethal force	-7.425	6.832	0.277	-1.354	5.340	0.800
Discretionary arrests	-4.542	6.709	0.498	-3.447	4.737	0.467
Officer injuries (days off)	-0.460	0.299	0.125	-0.106	0.248	0.668
Officer activities (index)	-0.008	0.025	0.743	0.028	0.029	0.320

Panel B: Effects on Field Outcomes by Crime Rate

	Sit-D			**Sit-D × Crime Rate**		
	Coef (1)	SE (2)	p-value (3)	Coef (4)	SE (5)	p-value (6)
Uses of non-lethal force	-10.656	8.249	0.197	0.171	0.723	0.813
Discretionary arrests	-5.378	7.936	0.498	-0.284	0.642	0.658
Officer injuries (days off)	-0.521	0.370	0.159	-0.005	0.034	0.882
Officer activities (index)	0.001	0.029	0.980	0.002	0.003	0.522

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 (4). 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, 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 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.

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, the year before 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.[28] 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

[28] 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 comprising multiple 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 \$1062 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 (\$1062) exceeds the cost of Sit-D inclusive of the refresher training. This underscores the promise of SitD 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	-	2.3 days
Personnel Costs Saved per Officer Trained	-	\$1,062

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.