NBER WORKING PAPER SERIES

POLICING AND MANAGEMENT

Max Kapustin
Terrence Neumann
Jens Ludwig

Working Paper 29851
http://www.nber.org/papers/w29851

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2022

The University of Chicago Crime Lab is an independent, non-partisan academic research center founded in 2008 to help cities identify the most effective and humane ways to reduce gun violence and reduce the harms associated with the administration of criminal justice. We thank the Chicago Police Department for making available the data upon which much of this analysis is based. The Chicago Police Department reviewed this publication for the limited purpose of ensuring personally identifying information was appropriately protected. We thank the City of Chicago, the Institute for Research on Poverty at the University of Wisconsin-Madison and Ken Griffin for financial support of this work, and AbbVie, the Joyce Foundation, the John D. and Catherine T. MacArthur Foundation, the McCormick Foundation, and the Pritzker Foundation for their support of the University of Chicago Crime Lab and Urban Labs, as well as Susan and Tom Dunn and Ira Handler. We thank Sydney Eisenberg, Rowan Gledhill, Katie Larsen, Riddhima Mishra, Michael Ridgway, and Noah Sebek for assistance with the data analysis, thank Roseanna Ander, Sean Malinowski, Marjolijn Bruggeling-Joyce, Anthony Berglund, Heather Bland, Trayvon Braxton, Amanda Dion, Mariah Farbo, Noe Flores, Jaureese Gaines, Brendan Hall, Alexander Heaton, David Leitson, Kevin Magnan, Emma Marsano, Jacob Miller, Ashley Orosz, Paulina Pogorzelski, Daniel Rosenbaum, Zoe Russek, Thomas Scholten, Kimberley Smith, Lauren Speigel, Diamond Thompson, Michael Thompson, Matthew Triano, and Yida Wang for invaluable assistance with the project more generally, and thank Nikolay Doudchenko, Michael Robbins, and Kaspar Wüthrich for sharing helpful code. For helpful comments we thank Aaron Chalfin, Philip Cook, John Donohue, Oeindrila Dube, William Evans, Barry Friedman, Elizabeth Glazer, Jeffrey Grogger, Candice Jones, Tracie Keesee, Christy Lopez, Justin McCrary, Sendhil Mullainathan, Daniel Nagin, Emily Owens, Wesley Skogan, Chad Syverson, and seminar participants at theAssociation for Public Policy Analysis and Management, Cornell, National Bureau of Economic Research, National Institute of Statistical Sciences, New York University, the University of California Los Angeles, University of Pennsylvania, and the University of Wisconsin-Madison. All opinions and any errors are our own and do not necessarily reflect those of our funders or of any government agencies. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Policing and Management
Max Kapustin, Terrence Neumann, and Jens Ludwig
NBER Working Paper No. 29851
March 2022
JEL No. H41,H75,J0,M0

ABSTRACT

How can we get more 'output,' and of the right sort, from policing? The question has only taken on greater importance with recent, widely publicized instances of police misconduct; declines in public trust in police; and a rise in gun violence, all disproportionately concentrated in economically disadvantaged communities of color. Research typically focuses on two levers: (1) police resources, and (2) policing strategies or policies, historically focused on crime control but increasingly also on accountability, transparency, and fairness. Here we examine a third lever: management quality. We present three types of evidence. First, we show there is substantial variability in violent crime and police use of force both across cities and within a city across police districts, and that this variation is related to the timing of police leader tenures. Second, we show that an effort to change police management in selected districts in Chicago generates sizable changes in policing outcomes. Third, as part of that management intervention the department adopted a predictive policing tool that randomizes which high-crime areas it shows to officers. We use that randomization to generate district-specific measures of implementation fidelity and show that, even within the context of a management intervention designed to improve implementation of the department's strategies, there is variability in implementation.

Max Kapustin
Department of Economics
Brooks School of Public Policy
Cornell University
Martha Van Rensselaer Hall
Ithaca, NY 14853
kapustin@cornell.edu

Jens Ludwig
Harris School of Public Policy
University of Chicago
1307 East 60th Street
Chicago, IL 60637
and NBER
jludwig@uchicago.edu

Terrence Neumann
University of Texas, Austin
Terrence.Neumann@mccombs.utexas.edu

A data appendix is available at http://www.nber.org/data-appendix/w29851

1 Introduction

How can we get more 'output,' and more 'outputs' of the right sort, from policing? The question has only taken on greater urgency on the heels of the murder of George Floyd in 2020, which was followed by marches across the United States and other countries demanding change, as well as a growing body of evidence documenting racial bias in policing and other criminal justice decisions (see, e.g., Arnold et al. (2020); Fryer (2020); Goncalves and Mello (2021); Hoekstra and Sloan (2020)). Public trust in the police has been declining,[1] particularly in communities of color, while the burden of gun violence has increased, disproportionately so in these same communities.

Research on policing has traditionally focused mostly on two types of levers. The first is police resources, more of which seem to reduce crime, particularly violent crime, and to simultaneously reduce arrests for serious crimes. At the same time more police seem to increase arrests for minor offenses, particularly in minority communities.[2] The second lever is policing strategies or policies, historically focused mostly on crime control but increasingly also on accountability, transparency, fairness, and legitimacy.[3] 'Blue-ribbon' committees convened by the National Academy of Sciences conclude that a department's choice of strategies and policies can also matter for various policing outcomes (e.g., National Research Council, 2004; National Academies of Sciences, Engineering, and Medicine, 2018).

Yet there seems to be more variation across cities in policing outcomes than just these two levers alone can explain. Consider, for example, what has happened over the past century in the three largest U.S. cities: Chicago, Los Angeles, and New York. Murder rates per capita in these cities tracked closely for most of the past 100 years, reaching similar peaks at the height of the crack cocaine epidemic in the early 1990s (Figure 1).[4] But they have diverged dramatically the

[1] https://councilonj.org/public-perceptions-of-the-police/

[2] See, e.g., Levitt (1997, 2002); Di Tella and Schargrodsky (2004); Klick and Tabarrok (2005); Evans and Owens (2007); Draca et al. (2011); Machin and Marie (2011); Owens (2013); Chalfin and McCrary (2018); Mello (2019); Durlauf and Nagin (2011); Chalfin et al. (2020).

[3] See, e.g., Tyler (2003); Meares (2008); Harcourt (2009); Meares et al. (2015); Owens et al. (2018); Bell (2021).

[4] One notable exception to this trend is the Prohibition era of the 1920s, when gang violence made infamous by Al Capone caused Chicago to pull away from its peers.

past 30 years. Murder rates fell 80-90% in NYC and LA, while Chicago's murder rate today is nearly back to what it was in the early 1990s and is actually higher now in predominantly African-American neighborhoods. While this divergence could be due in principle to such 'root causes' as socioeconomic conditions, poverty trends in the past 30 years across these cities are not notably different.[5] Criminologists instead often tell the stories of NYC and LA as being due in large part to policing changes (e.g., Stone et al., 2009; Zimring, 2011). This divergence does not appear to be explained by greater police resources in NYC or LA versus Chicago (Figure 2).[6] Nor do there seem to be major differences in policing strategies; these three departments, as is the case in most U.S. cities, claim to follow the same basic 'playbook' focused on high-crime people, high-crime places, and community policing (Table 1).[7] Something else is going on as well; what is it?

This paper explores a third lever that has been under-appreciated in both the scholarly literature and public discussion on policing: not what departments *aim* to do—or the resources available to them—but *how* they do it, which we refer to interchangeably as 'management quality' or 'implementation quality.'[8] Economic studies of the *private* sector view management as a 'technology' or 'intangible capital' that can increase productivity holding inputs and goals constant, can stem from either management practices or manager skill, and can be endogenously changed (Syverson, 2011; Bloom et al., 2016). An alternative view from organizational economics is that management practices vary because firms optimally adapt to local conditions, implying there is no

[5] Of the three cities, Chicago experienced the largest reductions in poverty rates between 1990 and 2019, from 21.6% to 16.4%. New York City's poverty rate fell from 19.2% to 16% during this period, while in Los Angeles the poverty rate increased from 16% to 16.7%. (Data for 2019 come from the American Community Survey; data for 1990 for Chicago from https://datahub.cmap.illinois.gov/dataset/9e0f8479-3dc7-4715-8711-be138f3bfb83/resource/5714aac0-9da1-4979-8568-c546d8adb771/download/nipcdatabul9611990censusselectedpovchar.pdf, for New York City from https://furmancenter.org/files/sotc/SOC_2016_FOCUS_Poverty_in_NYC.pdf, and for Los Angeles from https://laane.org/wp-content/uploads/2019/04/TheOtherLosAngeles_es.pdf.)

[6] While Figure 2 focuses on sworn officers per capita, the picture is qualitatively similar when considering all police employees, including civilians (Appendix Figure 1).

[7] See also National Academies of Sciences, Engineering, and Medicine (2018).

[8] One exception to this may be the small literature on the effects of police suddenly *not implementing* their strategy during work slowdowns (Mas, 2006; Chandrasekher, 2016, 2017; Sullivan and O'Keeffe, 2017). The findings of this literature suggest that such slowdowns, which typically only affect low-level enforcement activity like writing tickets or issuing summonses, may increase officer misconduct but have little influence on crime, particularly serious violent crime. That conclusion is consistent with findings that heightened enforcement of low-level offenses in New York City as part of 'broken windows policing' did not reduce violent crime (Harcourt and Ludwig, 2006, 2007; Harcourt, 2009).

universally better or worse set of practices (Gibbons and Roberts, 2012). Yet the data show substantial variation in private sector firms' output, much of which seems to be explained by differences in management practices.[9] This literature also shows that firms facing less competition are less productive and well-managed on average, and that improving a firm's management practices can raise its productivity.

The importance of management in the private sector raises the natural hypothesis that management might be important for the *public* sector as well—including policing. By police management we mean not just whether leaders follow best practices like collecting data to inform, evaluate, and adjust decisions. The role of police management runs deeper. If, within existing institutional constraints, the wrong officers are hired, or assigned or promoted to the wrong jobs, that is a failure of management. If front-line officers do not know what to do or how to do it, or choose not to do it, or police in biased ways, that is a failure of managing systems for training, planning, communicating, and accountability. There is no shortage of reasons to believe management quality might vary across places and over time. Accountability for public sector agencies comes from voters, who only indirectly influence policing since a vote for mayor is a choice of a *bundle* of policies. There are also low-information voters,[10] and even well-informed voters will struggle with the inference problem of isolating how much police versus other factors contribute to social conditions (Wolfers, 2002).

It is not hard to see examples of differences in police management in practice. Consider CompStat, an accountability tool used across the country in which local police commanders ('middle management') appear regularly before leadership to report on crime patterns and their plans to address them (Weisburg et al., 2008). Observing CompStat in New York, the department's most senior leadership asks detailed questions of precinct-level staff who must be prepared to provide

[9] For example, total factor productivity (TFP), or output from the same set of inputs, at the 90th percentile of U.S. manufacturing firms is *twice* as high as at the 10th percentile (Syverson, 2011). Similar variation has been documented in other countries as well (Bartelsman et al., 2013; Hsieh and Klenow, 2009). Bloom et al. (2016) find management variation accounts for around a third of cross-country TFP differences with the U.S. and about 30% of the 90-10 difference in TFP within countries. See also, e.g., Bloom and Van Reenen (2007, 2010); Bloom et al. (2013, 2017).

[10] For example, public perceptions of crime trends notoriously diverge from actual trends (https://www.pewresearch.org/fact-tank/2016/11/16/voters-perceptions-of-crime-continue-to-conflict-with-reality/).

detailed updates about ongoing cases and operations. In Chicago the past several decades, this level of preparation and detail has been more sporadic, as has been attendance by top department leadership or even the frequency with which CompStat was held at all. We see similar variation in what happens at roll call and other practices across departments.

Measuring police management practices and their consequences is challenging partly for want of reliable data. For example, Garicano and Heaton (2010) find that information technology (IT) investments have no detectable effects on crime on average, and also present suggestive evidence that complementarities with management practices like CompStat might enhance IT's effects. But the study relies on the Law Enforcement Management and Administrative Statistics (LEMAS) survey for data on management practices, which, as the authors note, only captured CompStat-related management practices towards the end of their panel and reveals some inconsistencies in departments' responses over time, possibly due to limitations of the survey questions. Canales et al. (2020) tries to solve the data insufficiency problem by surveying Mexican police departments about their managerial practices and shows those are related to employee turnover.[11]

We take a different approach here, presenting three complementary types of evidence about whether management varies for policing in the U.S. context, and with what consequences. First, we show there is substantial variation in policing outcomes in a panel of large departments. Even after controlling for city and time fixed effects and the number of sworn officers, there remains significant variation in murder rates per capita.[12] We show even greater variability for an outcome presumably even more directly under police department control: civilians killed by police. In principle we would also wish to examine broader measures of the collateral consequences of

[11] While understanding the productivity of public sector agencies has been the focus of decades of research (e.g., Wilson, 1989; Lynn, 2011), empirical work has been complicated by the difficulty of drawing inferences from agency-wide management changes over time or comparisons across agencies. As a result, as Lynn (1987, p. 187) argued, "knowledge in this field will always and necessarily be more conjectural and intuitive than normal social science." Among the modest exceptions to that forecast is a small literature in education that seeks to estimate principal value-added; see, e.g., Branch et al. (2009); Clark et al. (2009); Grissom and Bartanen (2019). There is a larger recent literature estimating value-added or other measures of productivity or problematic behavior such as racial bias by front-line public sector employees such as teachers (Jackson et al., 2014), prosecutors (Abrams and Yoon, 2007), judges (Abrams et al., 2012), and police officers (Goncalves and Mello, 2021).

[12] We focus on murder because it is the most accurately measured crime, and one that is available for most cities and time periods in recent years. Murder also overwhelmingly drives the total social harms from crime in U.S. cities; see, e.g., Chalfin and McCrary (2018).

policing, such as public sentiment towards the police, but unfortunately such measures are not readily available for very many cities at different points in time (e.g., Bell, 2021).[13]

To help isolate the role of police management, we show variation in outcomes across cities aligns with the tenures of departments' leaders, using the randomization inference procedure from Berry and Fowler (2021). We show similar results for a within-city, across-district analysis in Chicago. Variability in outcomes could in principle reflect variability in local preferences, not management quality, if there is a trade-off to some degree between crime prevention and police use of force (e.g., as a consequence of more aggressive policing). But we show that in the two-dimensional space defined by leader-specific fixed effects on violent crime and police killings of civilians, some leaders appear to dominate others (fewer violent crimes *and* enforcement harms). This correlational evidence suggests a combination of managerial skill and specific managerial practices could be important in explaining variation in policing outcomes.

Second, we examine the *causal* question of whether it is possible to push police departments closer to their production possibility frontiers (PPFs) through an intervention in Chicago intended to improve district-level management, adopted in response to a surge of gun violence in the city in 2016. The specific changes we examine, known as Strategic Decision Support Centers (SDSCs), were implemented with the help of the Bureau of Justice Assistance of the U.S. Department of Justice (DOJ) under the Obama Administration through a partnership between the Chicago Police Department (CPD) and the chief of staff of the Los Angeles Police Department (LAPD) at the time (Sean Malinowski). Our research center (the University of Chicago Crime Lab) provided analytical support to the SDSCs until the city could eventually hire its own crime analysts (see Appendix C for details and a discussion of potential conflicts of interest). The SDSCs tried to

[13] Developing reliable, low-cost, high-frequency measures of public sentiment available at sufficiently detailed geographic resolution should be a top priority for future research. Arguably one of the leading police departments in the world to date in measuring public sentiment is the Metropolitan Police Department in London, which surveys 12,800 residents each year. But the use of traditional surveys is expensive, which means that these community reports are available only quarterly, and are intended to be representative at only fairly large geographies (see, e.g., https://www.london.gov.uk/what-we-do/mayors-office-policing-and-crime-mopac/data-and-statistics/taking-part-mopacs-surveys). The lack of such measures means that police departments can hold middle managers accountable for measures of reported crime or arrests at CompStat meetings, but not for public sentiment.

improve day-to-day planning and management, and increased the information available to district commanders through a crime analyst and additional technology like cameras, recognizing people hold different normative views about such technology.[14] The SDSCs did *not* include a change to policing strategies or more officers, whose salaries and pension costs account for 90% of CPD's budget; the SDSC start-up costs were < 3% of each district's share of the CPD budget.[15] The SDSCs also did not involve systematically changing district commanders, and so helps us distinguish the effects of manager skill versus specific management practices.

These management changes were prioritized for those districts with the highest levels and largest recent increases in gun violence, which were also among the city's lowest-income, predominantly African-American neighborhoods. This prioritization complicates efforts to construct adequate statistical comparison groups, a challenge compounded by the fact that the SDSCs' introduction in early 2017 came on the heels of a 60% surge in homicides in 2016. We combine the greater variation in rates of violence and police use of force across individual police *beats* with an array of panel data estimators—ranging from difference-in-differences to synthetic controls (Abadie et al., 2010) to newer methods (Doudchenko and Imbens, 2017; Ben-Michael et al., 2021)—to construct comparison groups that most reliably estimate counterfactual outcomes in pre-SDSC data.[16] We also modify the standard permutation-based inference procedure of synthetic controls by creating artificial donor districts through a bootstrap re-sampling procedure, allowing us to overcome the limitations created by there being only 22 police districts in Chicago.

Our estimates suggest the SDSCs pushed districts toward their PPFs through at least the first three months of adoption: shootings and violent felonies declined by 32% the first month and 21% through three months, without statistically detectable changes in arrests, stops, or uses of

[14] For example, one 2007 survey of American adults found 71% supported increased use of cameras in public places. Support was lower for some groups, such as young people (18-29), at 61%, and Black Americans, at 63%: https://abcnews.go.com/images/US/1041a5Surveillance.pdf. We have been unable to find similar survey data from Chicago specifically on public support for cameras or other police technology, like gunshot detection systems.

[15] https://www.civicfed.org/civic-federation/blog/what-chicago-police-department-budget

[16] The one other effort of which we are aware to estimate the SDSCs' impacts relies solely on a two-way fixed effects difference-in-differences estimator (Hollywood et al., 2019). In addition to being susceptible to the measurement challenges described above, recent work has shown that, in settings such as this one with staggered treatment adoption and likely heterogeneous treatment effects, two-way fixed effects difference-in-differences estimates can exhibit substantial bias (Goodman-Bacon, 2021).

force. These impacts are sizable, large enough to reduce the Black-White disparity in gun-violence victimization rates citywide in Chicago by 13% if they persisted. However, those sizable initial reductions in gun violence appear to attenuate beyond three months, which may or may not be related to changes in the composition over time in arrests (with an increase in the share of arrests for drug and gun charges, but no clear change in the total overall number of arrests). These patterns reflect the difficulty of sustaining management changes, which itself reinforces the importance and variability of management practices.

Our final empirical analysis shows that even for an intervention like the SDSCs designed to improve the fidelity of implementation to CPD's policing 'playbook,' we still see variability in the implementation of that intervention itself. We then ask whether that simply reflects optimal adaptation of police managers to their local contexts, as suggested by the "management as design" view in organizational economics (Gibbons and Roberts, 2012; Bloom et al., 2016). We focus specifically on one part of the SDSC changes: the use of a predictive policing tool (HunchLab) that identifies small geographic areas or 'boxes' in which to prioritize additional police patrol resources. Our goal is not to determine whether predictive policing tools are social welfare maximizing or not, which is beyond our scope here. HunchLab adoption was a City decision, not one made by our research team, and we recognize there are important open questions about these tools.[17] Our goal instead is to examine how this tool was (or was not) used as a way to understand better the role of management variation specifically.

The most relevant feature of HunchLab for our purposes is that which boxes are shown to officers each shift is partially random. The large number of HunchLab crime hot-spot 'lotteries' at the district, day, shift, and local-area levels lets us generate district-specific estimates for the effects of HunchLab on how police spend their time. We see substantial variation across districts in the

[17] Given the high levels of racial residential segregation in Chicago, the use of any predictive policing tool to allocate police resources *across* neighborhoods has great scope for generating racial disparities in police contacts or the harms from enforcement. But with the SDSCs, HunchLab was used to allocate police resources *within* Chicago's 277 police beats; given the city's segregation, these beats are overwhelmingly comprised of residents from just a single racial or ethnic group. For a discussion of the larger concerns with predictive policing, see, e.g., Ferguson (2016); Shapiro (2017), and for evidence from a randomized trial carried out in Los Angeles, see Mohler et al. (2015); Brantingham et al. (2018).

degree to which officers spend more time in HunchLab-flagged boxes. We find no clear evidence to support the 'contingent management' hypothesis that commanders have private information about the differential effects of officer time in different areas across districts, and so are optimizing based on different local conditions. Specifically, there does not seem to be a systematic relationship between the district-level 'first stage' (effect of HunchLab on officer time spent in shown boxes) and 'second stage' (effect of patrol time in the boxes on shootings), although these estimates are somewhat noisy. Put differently, it does not appear that the districts where HunchLab's suggestions are ignored are the districts where additional officer time is least productive at reducing shootings; some districts seem to have just not followed the playbook.

Our hope is to help stimulate more research on police management and its determinants and consequences, which can have important policy as well as scientific value. Learning more about how to improve police management towards stated police objectives, which we take as given here (without a normative view about what these policies *should* be), will become only more important over time as the public and its elected representatives begin to think more about what policing should look like in the future and set new goals for their departments.

2 Management Variation Across- and Within Cities

2.1 Cross-city evidence: variation in police outputs

Though most large police departments in the U.S. adopt similar strategies (Table 1), how these strategies are implemented seem to vary both across departments and within departments over time. Take the recent history of changes to police management practices in Chicago. Under a succession of department superintendents, CPD introduced CompStat in 2003, suspended it in 2008, reinstated it in 2011, suspended it again in 2016, reinstated it again in 2017, and completely overhauled it in 2019.[18] A major change to how the department promotes officers made by an

[18] https://www.chicagotribune.com/news/ct-xpm-2011-04-30-ct-met-mccarthy-police-chief-20110430-story.html, https://www.policemag.com/344044/chicago-police-chief-suspends-weekly-compstat-meetings, https://www.chicagomag.com/city-life/october-2012/garry-mccarthys-new-chicago-crime-

interim superintendent was promptly reversed by his successor.[19]

To understand how variability in police management across and within cities might give rise to variability in policing outcomes, we start by documenting large residual variation across departments in two policing outcomes that have particularly important social consequences: homicides and civilians killed by police. While there are obviously many additional policing outcomes that are of social importance, such as public perceptions of the police or subjective well-being more generally, there is no systematic collection of such measures across cities over time. Aside from arrests, the government does not even collect consistent measures across cities of different police activities, such as traffic or pedestrian stops.[20]

For practical reasons we focus here on the 50 departments serving the largest jurisdictions in the U.S.[21] For the period 2010-2019 we measure the homicide rate using the FBI's Uniform Crime Reporting (UCR) program (Kaplan, 2020) and the rate of civilians killed by police using data collected by Fatal Encounters.[22] The UCR data also include the number of sworn officers in each department and year. We use this information to estimate population-weighted two-way fixed effects models of the form:

$$Y_{dt} = \beta X_{dt} + \gamma_d + \lambda_t + \varepsilon_{dt} \tag{1}$$

where Y_{dt} is a measure of police output for department d in year t, X_{dt} is the rate of sworn officers per capita, and γ_d and λ_t are department and year fixed effects, respectively, which help account for variation across cities and over time in socio-demographic factors and other determinants of

strategy-social-networks-hot-people/, https://www.nydailynews.com/news/crime/ex-chicago-top-blames-city-spike-violence-politicians-article-1.2925908.

[19] https://www.npr.org/local/309/2019/12/11/787040792/beck-suspends-controversial-merit-promotions-in-police-department, https://www.chicagotribune.com/news/criminal-justice/ct-chicago-police-david-brown-decisions-merit-promotions-20210722-svq5ac4qnjdkrhna6bfhbc6wki-story.html.

[20] Non-governmental organizations like the the Stanford Open Policing Project have tried to fill the gap, but they capture a convenience sample of departments from which data could be obtained and for inconsistent time periods across the different agencies; see https://openpolicing.stanford.edu/data/.

[21] These 50 jurisdictions range in population from just over 460K (Mesa, AZ) to 8.4M (New York, NY) and include nearly 55M people in total, or almost a fifth of the U.S. population. We refer throughout this section to these 50 jurisdictions as "cities," although 10 are counties.

[22] We remove a small number of incidents from the Fatal Encounters dataset where police officers killed civilians while off-duty.

crime and other outcomes. The residuals from this regression represent variation in an outcome in a given year not explained by time-invariant differences across departments, shocks common to all departments each period, or sworn officers.

Even after controlling for department and year fixed effects, as well as the number of sworn officers, there remains substantial variability in both outcomes (Figure 3).[23] The standard deviation of the homicide rate residuals (2.6 per 100,000) is over a quarter of the mean homicide rate (9.6 per 100,000). Since many time-varying, city-specific factors outside the control of police departments may affect homicide rates, it is noteworthy that we also see residual variation in police killings of civilians, over which departments likely have more direct influence. The standard deviation of these residuals (0.26 per 100,000) is nearly three quarters as large as the mean rate of civilians killed by police (0.37 per 100,000). Notably, the residual variation in both of these measures is greater for Chicago than it is for either New York or Los Angeles.

Could variation in police management help explain at least part of this variation? To answer this we incorporate data on the exact timing of changes to police leadership and estimate a permutation-based procedure known as randomization inference for leader effects, or RIFLE (Berry and Fowler, 2021). RIFLE works by measuring how much leaders' tenures explain outcome variability and then comparing this to how much a placebo set of tenures can explain the same variability. If leaders affect an outcome, then their real tenures are likelier to explain more of the outcome's variability than a placebo set of tenures.

To obtain this placebo set of tenures, we permute (shuffle) leaders' tenures within a department, keeping the length of each leader's tenure the same as in the real data but changing the order in which they served. For this permutation procedure to produce placebo tenure sets that differ from leaders' actual tenures, there must be variation in the lengths of leaders' tenures within a department. For example, if we have 4 years' worth of data for a department split between 2 separate leaders' tenures of 2 years each, permuting the order of these tenures has no room

[23] After controlling for department and year fixed effects, the UCR measure of sworn officers has virtually no explanatory power for either outcome. This may be partly due to the significant measurement error in that variable (e.g., Chalfin and McCrary, 2018).

to explain the variation in outcomes differently. But suppose we had 3 years of data from a department in which the first leader is in charge for the first 2 years and the second leader is in charge for just the third year; intuitively, the permutation test is asking whether the second year's outcome is more similar to the first year's outcome than to the third year's.

We apply RIFLE to department-by-month data on rates of homicide, violent index crimes,[24] all arrests, and narcotics arrests from the UCR, and data on civilians killed by police from Fatal Encounters, for the same group of 50 departments from 2010 to 2019. For these departments, we hand-collected public information on the tenures of their police chiefs.[25] Before implementing the RIFLE procedure, we first convert each outcome Y in department d under police chief c in month t into a z-score by subtracting from it the department-specific mean and dividing it by the department-specific standard deviation to account for the large differences across departments in rates of crime, enforcement activity, and enforcement harm, and therefore in the magnitude of month-to-month variability as well.[26]

$$z_{dct} = \frac{Y_{dct} - \bar{Y}_d}{\text{sd}(Y_{dct})}$$

Then, we residualize these z-scores to remove calendar-month time effects common to all departments in our sample:

$$\hat{e}_{dct} = z_{dct} - \hat{z}_{dct} = z_{dct} - \sum_{r=1}^{T} \hat{\delta}_r \mathbb{1}[t = r]$$

The RIFLE procedure then proceeds in four steps:

1. Store the observed $R^2 = R^2_{obs}$ from a regression of the outcome (residualized z-score) on leader tenure fixed effects:

$$\hat{e}_{dct} = \gamma_{dc} + \varepsilon_{dct}$$

[24] Violent index crimes include aggravated assault, forcible rape, murder, and robbery.

[25] The earliest reliable information we could locate about the tenures of police chiefs begins in the late 1990s and early 2000s for most of the 50 departments in our set.

[26] Our results are qualitatively similar when not first normalizing each observation by converting it into a z-score.

2. Randomly permute the order of leaders' tenures within each department, shuffling each tenure as a block.

3. Re-run the same fixed effects regression as in (1), storing that iteration's $R^2 = R^2_{(i)}$.

4. Repeat steps (2)-(3) for $i = 1, ..., N$ iterations, where $N = 5,000$.

After completing steps (1)-(4), we calculate the p-value for an outcome as:

$$p = \frac{\sum_1^{50} \mathbb{1}\left[R^2_{(i)} > R^2_{obs}\right]}{50} \tag{2}$$

We report our results in Table 2. At a 5% significance level, we can reject the null hypothesis of no leadership effects for violent index crimes and civilian-police killings. We fail to reject the null for all arrests, narcotics arrests, while the p-value for homicides is $p = 0.17$.

That police chiefs appear to affect some of the most socially costly police-related outcomes—violent index crimes and civilian-police killings—by itself need not be due to differences in management quality. In principle there could be trade-offs between policing activities that reduce crime and those that minimize the harms from law enforcement (see, e.g., the discussion in Devi and Fryer (2020)). If that were true, then different chiefs (or their constituents) could simply have different preferences about how to make that trade-off. For example, some forms of proactive or interventionist policing may be effective at reducing crime, but may also result in more interactions with the public that could result in harm. Distinguishing between the "trade-offs/preferences" and "dominance/skills" hypotheses requires more than just documenting the existence of leadership effects; we need to understand something about the structure or covariance of leadership effects on different outcomes.

We look for evidence to help distinguish between the "trade-offs/preferences" and "dominance/skills" hypotheses by estimating leader tenure fixed effects for violent index crimes and civilian-police killings, and comparing these estimates in two-dimensional outcome space. Specif-

ically, we estimate population-weighted equations of the form:

$$p_{dct} = \gamma_{dc} + \varepsilon_{dct} \tag{3}$$

where $p_{dct} = \frac{Y_{dct}}{\bar{Y}_d} - 1$ is the percent deviation of the outcome Y in department d under chief c in month t from the department mean. The parameters of interest are the leader tenure fixed effects, γ_{dc}. These fixed effects, though not necessarily the causal effects of each chief, represent the average percent deviation in an outcome from the department mean during a chief's tenure. Because these outcomes are often noisy, we also replicate our results by calculating value-added (VA) effects for each chief using the shrinkage estimator in Easterly and Pennings (2020).[27]

Figure 4 reports leader tenure fixed effects estimates for violent index crimes and civilian-police killings, for all tenures of at least 6 months.[28] A quarter of the tenures shown have average deviations from their department mean of at least 16% for violent index crimes and 36% for civilian killings by police. Tenures on the diagonal running from the upper left to the bottom right quadrants are those with outcomes consistent with the "trade-offs/preferences" hypothesis, having lower rates of violent crime but higher rates of enforcement harms (upper left), or higher rates of violent crime but lower rates of enforcement harms (bottom right), relative to their peers. Tenures on the diagonal running from the bottom left to the upper right are those with outcomes consistent with the "dominance/skills" hypothesis, having higher rates of both violent crime and enforcement harms (upper right), or lower rates of both outcomes (bottom left), relative to their peers. This evidence is not consistent with the idea that "trade-offs/preferences" explains all of the variability: there are as many tenures in the upper left and bottom right quadrants (115) as the number in the upper right and bottom left quadrants (113).[29] The shrinkage estimator from

[27] The VA effects estimates, like the fixed effects estimates, are not causal, as we do not exploit any exogenous variation in the assignment of chiefs to departments. In Monte Carlo simulations to recover leader effects on countries' growth rates, Easterly and Pennings (2020) find that VA effects from a shrinkage estimator are both more efficient than fixed effects estimates and forecast unbiased.

[28] The results are qualitatively similar when including all tenures, though the fixed effects estimates for shorter tenures are often outliers.

[29] If we limit the analysis to just those tenures where we can reject the null of one or the other of a tenure's fixed effects being zero (at a 10% significance level), then 90 and 87 tenures, respectively, fall in the "trade-offs/preferences" and

Easterly and Pennings (2020) yields similar results (Appendix Figure 2).

This suggests the potential importance of management quality. However, the results are limited in two important ways. First, in both the two-way fixed effects and RIFLE analyses, there remains a large scope for unmeasured differences across and within departments over time unrelated to management quality that could affect the outcomes we measure. Second, the set of policing outcomes consistently measured over time for 50 departments is unfortunately quite limited. For these reasons, we turn next to an analysis of variation in a broader set of outcomes across police districts *within* a single city: Chicago.

2.2 Within-city evidence: commander leadership effects

Chicago is divided into 22 police districts (Figure 5). While districts share the same department-wide policies and strategies, commanders have discretion in how to implement them. For example, commanders direct the day-to-day deployment of several hundred patrol officers in their districts, each roughly equivalent to a mid-sized city.[30] While commanders don't directly supervise patrol officers,[31] they supervise the supervisors, and may exert much greater influence over local patrol behavior than does leadership at headquarters (e.g., Mummolo, 2018).

We extend the cross-city analysis from section 2.1 to a within-Chicago, cross-district, over-time analysis. We started by hand-collecting public information on the tenures of Chicago's commanders, over a period of 12 to 17 years depending on the district,[32] and implement RIFLE, the results of which are reported in Table 3. At a 5% significance level, we can reject the null hypothesis of no leadership effects for shootings, stops, and uses of force, and at a 10% significance level we can reject the null for gun arrests. We fail to reject the null for broader categories of crime,

"dominance/skills" quadrants.

[30] A smaller district like the 15th, covering the West side neighborhood of Austin, is under four square miles and contains 60,000 residents. A larger district like the 8th, covering several South side neighborhoods including Chicago Lawn, is roughly 24 square miles and contains almost 250,000 residents, or about the same population as Buffalo, NY.

[31] Direct supervision of patrol officers is usually the responsibility of Sergeants and Lieutenants.

[32] Though most of the CPD outcome data are available from 2000 to 2020, the earliest reliable information we could locate about commanders' tenures ranges from May 2003 to March 2008.

like violent felony offenses (mostly armed robberies), as well as arrests.

These patterns are consistent with the intuition that outcomes largely at officers' discretion, such as street and traffic stops—and the gun arrests that sometimes result from them[33]—are most responsive to commanders' influence, while the outcomes more dependent on outside factors, like violent felony offenses, are less responsive. We might have expected officer discretion to also matter a great deal for overall arrests, and particularly misdemeanor and narcotics arrests. But the high p-values for these outcomes suggest that commanders either have little impact on how officers use their discretion to make most arrests, or there is little variation across commanders in how they try to direct this behavior among officers.

In Figure 6 are the results of estimating population-weighted commander fixed effects for shootings and uses of force for commanders with tenures of at least 6 months (equation 3), and identify their covariance by plotting both estimates in two-dimensional outcome space.[34] We do not find that variation in local preferences about policing (the "trade-offs/preferences" hypothesis) fully accounts for the variability we document in shootings and uses of force. A quarter of the commander tenures shown have average deviations from their district mean of at least 24% for shootings and 26% for uses of force. A total of 38 out of 111 commander tenures fall into the upper right and bottom left ("dominance/skills") quadrants. A shrinkage estimator from Easterly and Pennings (2020) produces similar results (Appendix Figure 3).

We might still be worried that where commander tenures fall in the two-dimensional outcome space of Figure 6 is determined by something other than commander effects. With the district-level data, unlike the city-level data, we can carry out one additional test by focusing on the 18 commanders who spent at least 6 months in charge of two separate districts ('switchers'). For each commander, we first calculate the distance in our two-dimensional outcome space (shootings

[33] Chicago police recovered 11,273 firearms in 2020 (https://www.nbcchicago.com/violence-in-chicago/where-police-recover-the-most-guns-in-chicago/2612202/); of these, 3,233 (29%) were recovered during street and traffic stops (https://www.chicagotribune.com/opinion/commentary/ct-opinion-data-points-gun-carrying-crime-lab-20210403-5iz6blr6urhlji7hxwyjwrnhc4-htmlstory.html). Such street and traffic stops often change in their frequency in response to changes in the level of what criminologists call 'pro-active policing.'

[34] As in the cross-city analysis, fixed effect estimates for shorter tenures are often outliers, but the overall results including them are qualitatively similar.

and uses of force, with both outcomes standardized into within-district z-scores) between their two tenures. Then we permute which two tenures belong to the same commander, recalculating each time the distances between each 'pseudo pair' of tenures, and measure where the distance between the actual pair of tenures falls within this permutation distribution. This is a low-power test, but the p-value of 0.2 is at least suggestive of the fixed effects estimates containing some signal about commander effects.

The data presented in this section both raise a puzzle—why there is so much variation in outputs across departments and across districts within a given department?—and provides evidence consistent with, although not by itself definitive proof for, one candidate explanation: variation in management. In what follows we subject this hypothesis to additional testing.

3 A Management Intervention: Chicago's SDSCs

The results in the previous section are consistent with the idea that departments (and districts) vary in their distance from their production possibility frontiers (PPFs). But these results are ultimately descriptive. We next try to measure the causal effects of an intentional, exogenous change in management quality by studying an intervention in Chicago that was deployed in some police districts and not others. We first describe the management changes, which would most clearly signal a move closer to its PPF if we observed at least some outcomes that are desirable from society's perspective "get better" and no outcomes "get worse."

The management changes we study were adopted in Chicago with the help of the Bureau of Justice Assistance of the U.S. Department of Justice (DOJ) under the Obama Administration, on the heels of a 58% increase in homicides from 2015 to 2016 in Chicago.[35] To advise on CPD's

[35] The causes of this increase remain unclear (Kapustin et al., 2017). Social conditions such as poverty and segregation, commonly cited as reasons for Chicago's heightened level of violence relative to its peers, did not change suddenly that year. Other hypotheses focus on changes that occurred immediately prior to the increase, such as the sharp decline in street stops conducted by CPD (Cassell and Fowles, 2018). However, in addition to conflicting evidence from the broader research literature about the relationship between street stops and violent crime (e.g., MacDonald et al., 2016; Weisburd et al., 2016; Rosenfeld and Fornango, 2017), the large number of additional changes involving CPD that occurred in a short period of time in late 2015—such as the firing of the Superintendent on the heels of the November 2015 release of a video showing CPD officer Jason Van Dyke shooting 17-year-old Laquan McDonald in

approach to crime reduction, DOJ invited to the city experts from other departments, including the LAPD, led by Chief Sean Malinowski.[36]

Working in partnership with the then-CPD Superintendent's senior management team, the outside experts identified shortcomings in CPD's implementation of its own policing strategy. That strategy, as in most major departments in the U.S. (Table 1), calls for a proactive focus on places and people at highest risk of violence, and for engaging residents through community policing.[37] However, in practice, the department's day-to-day operations often looked quite different:

- Despite collecting detailed data on reported crime and police activity (the CLEAR system), as well as developing a homegrown mapping software to analyze it (Caboodle), few front-line officers or their supervisors seemed trained in how to use these resources or actually used them, based on our observations and conversations with many CPD officers and supervisors.

- Roll calls occurring before each shift appeared to be largely cursory, with little in the way of specific instructions being communicated by supervisors to front-line officers or even data being used to direct them on where to focus their time. As one CPD commander put it, officers were "just patrolling randomly" and "riding around rubber-necking on the street waiting for something to happen."[38]

- CPD first began building its camera network in 2003, under then-Mayor Richard M. Daley.[39] By 2011, the police had access to 10,000 cameras, according to the ACLU.[40] Yet despite this enormous investment, relatively few officers outside of the department's single "fusion center" at headquarters appeared to utilize the cameras.

October 2014 and the subsequent launch of a DOJ civil rights investigation—make it impossible to determine how much any single change contributed to the violence increase.

[36] Also helping to lead the effort were Marjolijn Bruggeling and Terry Gainer.

[37] CPD was the first large department in the U.S. to embrace the community policing model with the establishment of the Chicago Alternative Policing Strategy (CAPS) in 1993 (Skogan and Hartnett, 1999).

[38] https://chicago.suntimes.com/2017/3/29/18358635/violent-crime-falls-in-2-districts-run-by-the-johnson-brothers

[39] https://home.chicagopolice.org/information/police-observation-device-pod-cameras/

[40] https://www.aclu-il.org/sites/default/files/field_documents/video_camera_surveillance_in_chicago.pdf

- CPD's community policing effort, CAPS, atrophied since its introduction in the mid-1990s. As a report by the CAPS evaluation team put it, "after opening with great anticipation, the program went stagnant and lacked true direction" (Skogan et al., 2002).

To remedy these limitations, the outside experts recommended the establishment of planning processes and hubs within each district, which came to be known as Strategic Decision Support Centers (SDSCs). The SDSCs sought to address the shortcomings in CPD's implementation at that time of its own policing strategy with respect to planning, information collection, and information analysis.

First, coinciding with the SDSCs' introduction, CPD introduced several technology enhancements that increased the amount of information available to commanders. This included an acoustic gunshot detection system (ShotSpotter); a place-based predictive policing software (HunchLab, described in section 5); an expansion of, and improved user interface for, CPD's existing network of Police Observation Device (POD) cameras;[41] and mobile phones for officers with access to both ShotSpotter and HunchLab. These technologies were intended to both aid officers in the field—such as by shortening response times after a shooting (ShotSpotter), giving guidance on where to focus additional patrol time (HunchLab), or providing an additional set of eyes on a location (POD cameras)—and to provide further input on local crime patterns for planning.

Second, the SDSCs tried to significantly increase the information made available to commanders by creating a role specifically to provide it: a civilian crime analyst. During the initial roll-out of the SDSCs, staff at our research center, the University of Chicago Crime Lab, served in the civilian crime analyst role until the City authorized the hiring of permanent analysts (for further details, see Appendix C). The analyst, who is trained on all of CPD's existing software tools like CLEAR and Caboodle, develops analytical products describing recent patterns of criminal activity in the district. An example is presented in Appendix Figure 4. In 2017, the commander of the 7th district asked the district's analyst to examine data on stolen vehicles, which are often used to

[41] Exact data on how many POD cameras were added or upgraded as part of the SDSC expansion is unavailable. According to Hollywood et al. (2019), as of early 2019, the total number of cameras to which CPD had access, including those owned by other agencies, approached 35,000. For an example of the improved interface, see Figure 1.5 in Hollywood et al. (2019).

commit shootings, to determine if there was a pattern. The analyst identified a cluster of 18 cars recently stolen from the adjacent district and all recovered near the same intersection. Based on this information, the commander ordered increased patrols that led to the arrest of a person with an extensive history of motor vehicle theft and a connection to an unsolved quadruple homicide.

Finally, consistent with findings from the private sector that changes in managerial practices are often important for enhancing the productivity of IT investments (Bloom et al., 2016), and similar suggestive evidence for policing (Garicano and Heaton, 2010), the SDSCs introduced a regular planning process for district commanders. This included a daily briefing for the commander to incorporate information supplied by the crime analyst into deployment plans. The briefing follows a standard format and includes information such as:

- recent crime trends and high-profile arrests
- high-priority open warrants (e.g., if someone is wanted for murder)
- deeper analyses into areas of interest, including those raised at previous briefings
- an overview of available discretionary resources and their current deployment locations

The output of the briefing is a set of missions ordered by the commander, along with information for dissemination to field units. Missions can vary in their complexity, ranging from dispersing trespassers at businesses associated with violence to heightened patrol activity for the anniversary of a slain gang member's death. The information produced by the SDSC is shared with officers during roll calls at the start of each watch (shift). For example, in one district, SDSC officers prepare a binder with the high-priority open warrants discussed in the briefing for tactical officers to review. The new planning process also included a daily assessment of the previous day's missions, and any adjustments that might be needed.

As important as understanding what the SDSCs were, it is equally important to understand what they were *not*:

- While there was a modest change in district-level resources (the crime analyst and some additional technology), SDSC districts did not receive an infusion of what is far and away

the main driver of CPD resources: officer time, either in the form of new officers being hired, existing officers being reallocated, or officers working additional overtime hours. For example, in the 7th district, one of the first to receive an SDSC, GPS data from police cars shows effectively no change in aggregate officer time before, during, or after the establishment of the SDSC (Appendix Figure 11).

- SDSC districts were not told to implement a new policing strategy. Rather, they were directed to utilize their SDSCs to "assist Department members with district-crime forecasting and achieving the primary mission of district crime-reduction."[42]

In other words, neither of the two levers that have been the focus of most policing research and policy were pulled in response to Chicago's violence spike; instead, CPD adopted an intervention that sought to strengthen the implementation of its existing efforts largely through improved district-level management.

4 Measuring the SDSCs' Impact

4.1 Empirical approach

To measure the SDSCs' impacts, we use the fact that their roll out across the city was staggered. The first two SDSCs launched in February 2017 in the 7th district (Englewood, on the South side) and the 11th district (Garfield Park, on the West side), which have historically had among the highest levels of violence in Chicago.[43] Six weeks later, following large declines in gun violence in both districts that were visible even in the raw data, the city launched SDSCs in the four remaining so-called *"Tier 1"* districts with elevated violence levels: the 6th, 9th, 10th, and 15th (Figure 5). We generate two complementary types of impact estimates:

[42] http://directives.chicagopolice.org/CPDSergeantsExam_2019/directives/data/a7a57b85-16c2efbe-c2416-c2fa-edbba6051837c01c.html

[43] The 7th and 11th districts have populations of around 30,000 and 37,000 residents, respectively; this is on the order of cities like Atlantic City, NJ, or Annapolis, MD.

- Impact estimates for each of the first two "early adopter" districts, the 7th and 11th, measured over the short-run period before the other Tier 1 districts adopted SDSCs as well. The advantage of this analysis is we are able to use data from the other high-violence Tier 1 districts as "donors" to construct comparison groups. The disadvantage is we are only able to look at a one-month follow-up period.

- Impact estimates for all of the Tier 1 districts as a whole. The drawback of this analysis is we are limited in constructing our comparison groups to using the Tier 2-4 districts that have lower overall average rates of violence, as we discuss below. The advantage is that we are able to estimate impacts for longer follow-up periods.[44]

Our goal is to estimate counterfactual outcomes using information from districts that had yet to receive an SDSC. Adopting the notation of Doudchenko and Imbens (2017), let $Y_{i,t}^{obs}$ be an outcome, like the rate of shootings per capita, observed in unit i in month t, where each unit is a geographic area such as a district. We have a panel of $J+1$ units observed for T months, where treated unit $i = 0$ receives an SDSC starting in month $t = T_0 + 1$ and donor units $i = 1, ..., J$ remain untreated for the entire panel. Using the potential outcomes framework (Rubin, 1974), we can express the observed outcomes as:

$$Y_{i,t}^{obs} = \begin{cases} Y_{i,t}(0) & \text{if } i > 0 \\ Y_{i,t}(0) & \text{if } i = 0 \text{ and } t \leq T_0 \\ Y_{i,t}(1) & \text{if } i = 0 \text{ and } t > T_0 \end{cases}$$

The parameter of interest is the average treatment effect, or the difference in outcomes for the treated unit due to the SDSC, for $t > T_0$:

$$\tau_0 = \frac{1}{T - T_0} \sum_{t=T_0+1}^{T} Y_{0,t}(1) - Y_{0,t}(0) \tag{4}$$

where $Y_{0,t}(1) = Y_{0,t}^{obs}$ is an observed outcome with the SDSC and $Y_{0,t}(0)$ is an unobserved potential

[44] We are unable to obtain reliable estimates for each of the other Tier 1 districts on their own, as we discuss in more detail in Appendix B.1.

outcome without the SDSC.

To estimate $Y_{0,t}(0)$ for $t > T_0$, we turn to several panel data estimators, most of which can be expressed as a linear combination of an intercept and the weighted sum of observed outcomes among donor units $i = 1, ..., J$:

$$\hat{Y}_{0,t}(0) = \hat{\mu} + \sum_{i=1}^{J} \hat{\omega}_i Y_{i,t}^{obs} \tag{5}$$

The panel data estimators we consider differ mainly in how they choose $\hat{\mu}$ and $\hat{\omega}_i$. For example, a standard difference-in-differences (DD) estimator allows for a fixed, non-zero $\hat{\mu}$ difference between the levels of the treated and donor units, and assigns equal, positive weights to all donor units that sum to one ($\hat{\omega}_i = \frac{1}{J}$). The synthetic control method (SCM) (Abadie and Gardeazabal, 2003; Abadie et al., 2010, 2015) does not admit an intercept ($\hat{\mu} = 0$) and chooses non-negative donor weights summing to one ($\hat{\omega}_i \geq 0, \sum_{i=1}^{J} \hat{\omega}_i = 1$) that minimize the covariate distance between the treated unit and the weighted sum of the donor units.[45] By restricting their donor weights to be non-negative and sum to one, the DD and SCM estimators avoid extrapolating outside the support of the data and therefore produce estimates that are less model-dependent (King and Zeng, 2006). In contrast, the elastic net (EN) estimator, introduced by Doudchenko and Imbens (2017), admits an intercept and uses regularized regression to limit the number of non-zero donor weights, but permits extrapolation by allowing those weights to be negative and their sum to exceed one.[46] The augmented synthetic control method (ASCM) of Ben-Michael et al. (2021) modifies traditional SCM by allowing donor weights to be negative in cases where the pre-treatment "fit" is poor.[47]

Based on our own experimentation with both simulated data and the pre-treatment SDSC data, we also consider a modified version of the EN estimator (EN-M) that restricts donor weights to be non-negative and still allows their sum to exceed one, but that uses a different cross-validation procedure to parameterize the elastic net penalty term. The original EN estimator

[45] In theory, the covariates can include both pre-treatment outcomes and auxiliary covariates. In practice, as Kaul et al. (2018) note, many researchers include the entire pre-treatment outcome path among the covariates, which renders any auxiliary covariates irrelevant.

[46] We implement the DD, SCM, and EN estimators using the `MCPanel` package in `R`, as described in Athey et al. (2018).

[47] We implement the ASCM estimator using the `augsynth` package in `R`.

chooses hyperparameters that minimize error in the post-treatment period among the donor units, under the assumption that, because they are untreated, their post-treatment period error should, on average, be zero. This procedure yields a single pair of hyperparameters to be used in determining weights for all the treated units, relying only on data from the donor units. This may be particularly problematic in our setting, where the SDSC districts are outliers relative to the donors (see, e.g., Appendix Figure 6). The hyperparameters minimizing post-treatment period error among those donor units may not yield the most reliable comparisons for the treated units.

Our modified version chooses hyperparameters that minimize error in the *pre-treatment period* for just the treated unit in question (e.g., the 7th or 11th district, or the Tier 1 aggregate). The assumption required by this modified approach (that treatment has not yet occurred in the pre-treatment period) is weaker than the one in the original Doudchenko and Imbens (2017) approach (that the donor units are unaffected by treatment in the post-treatment period). Furthermore, by choosing hyperparameters that minimize prediction error within the treated unit across a series of one-step-ahead forecasts, the estimator builds in some additional protection against overfitting (see Appendix A for details).

We face several challenges to estimating $Y_{0,t}(0)$ for $t > T_0$. For starters, shootings in Tier 1 districts, especially the 7th and 11th, have not only been among the highest per capita in the city in terms of *levels* but these districts also saw among the largest *increases* in 2016, the year preceding the SDSCs' launch (Appendix Figure 6). This creates a potential source of confounding if the SDSC districts were also likely to experience larger drops in gun violence in 2017 due to mean reversion. Guarding against this type of confounding requires finding a weighted set of donor units that experienced similar pre-treatment trends to the treated unit. But finding such a weighted set may be difficult when the treated unit falls outside the support of the donor pool and the estimator does not permit extrapolation. Even if these concerns are addressed, Abadie (2021) cautions against including units in the donor pool that differ substantially from the treated unit in attributes affecting the outcome of interest, as doing so can introduce bias. This may be particularly relevant in Chicago, where gun violence is concentrated in a handful of districts and

where patterns of violence may further differ between the city's South and West sides.[48]

We seek to overcome these challenges in two ways:

- First, we try expanding the donor pool so that the treated unit is less likely to fall outside its support. While Chicago only has 22 police districts, each district contains approximately a dozen smaller beats (277 in total across the city; see Figure 5), allowing us to take advantage of the much higher variation in outcomes like rates of shootings per capita across donor *beats* compared to donor *districts* (Appendix Figure 7).

- Second, we perform a data-driven backdating exercise suggested by Abadie (2021) that divides the treated unit's pre-treatment period in two, using T_0^{placebo} of the earliest months to construct a comparison and the later $T_0 - T_0^{\text{placebo}}$ months to assess its fit, for $T_0^{\text{placebo}} \geq \underline{T_0}^{\text{placebo}}$.[49] This procedure allows us to identify which estimator, donor pool, and donor type is likeliest to maximize out-of-sample prediction accuracy for each treated unit and outcome, relying solely on pre-treatment data. For additional details, see Appendix B.1.

To assess the statistical significance of our estimates, we use the permutation test suggested by Abadie et al. (2010). Given a choice of test statistic, the permutation test compares its value for the treated unit to a placebo distribution derived from permuting treatment status among the untreated districts.[50] The simplest choice of test statistic is the magnitude of the estimated average effect over the entire post-treatment period ($|\hat{\tau}_i|$). However, this may be large not because the true average effect has a large magnitude, but because the estimator does a poor job of estimating the counterfactual. Because an estimator's reliability may vary across units, Abadie et al. (2010)

[48] Specifically, as Hagedorn et al. (2019) note, a larger share of gun violence is thought to be the result of personal disputes on the South side than on the West side, where a larger share of shootings are thought to be related to the narcotics trade concentrated along the I-290 corridor (often called the "heroin highway").

[49] There are a total of $T_0 = 49$ months of pre-treatment data available, from January 2013 through January 2017. We discuss below and in Appendix B.1 the sensitivity of the backdating exercise to the choice of $\underline{T_0}^{\text{placebo}}$.

[50] Note that, while the estimation procedure can use a donor pool consisting of individual police beats rather than districts to solve the common support problem as outlined above, the inference procedure estimates placebo treatment effects only for individual police districts. Outcomes are often too rare in individual police beats to generate reliable counterfactuals—and therefore placebo treatment effects—for them.

suggest using the test statistic:

$$\theta_i = \frac{\left(\frac{1}{T-T_0}\sum_{t=T_0+1}^{T}(Y_{i,t} - \hat{Y}_{i,t})^2\right)^{1/2}}{\left(\frac{1}{T_0}\sum_{t=1}^{T_0}(Y_{i,t} - \hat{Y}_{i,t})^2\right)^{1/2}} \tag{6}$$

The denominator, a root mean squared prediction error (RMSPE), measures the magnitude of absolute deviation between a unit and its counterfactual across the pre-treatment periods. Dividing by the pre-treatment RMSPE penalizes units where the estimator's reliability is likely to be low, as judged by its ability to match a unit's observed pre-treatment outcome patterns. However, using post-treatment RMSPE for the numerator, as Abadie et al. (2010) suggest, results in units with large absolute deviations from the counterfactual in the post-treatment period having large values of the test statistic, even if the estimated average treatment effect over the post-period is small. If the post-treatment period is one time period in length, a large absolute deviation is synonymous with a large estimated treatment effect. But if the post-treatment period is more than one time period in length, it is possible for the estimated average treatment effect to be small even if the absolute deviation is large, such as if estimated treatment effects are of different signs across post-treatment periods. For this reason, we instead use the test statistic:

$$\theta_i = \frac{|\hat{\tau}_i|}{\left(\frac{1}{T_0}\sum_{t=1}^{T_0}(Y_{i,t} - \hat{Y}_{i,t})^2\right)^{1/2}} = \frac{\left|\frac{1}{T-T_0}\sum_{t=T_0+1}^{T} Y_{i,t}(1) - \hat{Y}_{i,t}(0)\right|}{\left(\frac{1}{T_0}\sum_{t=1}^{T_0}(Y_{i,t} - \hat{Y}_{i,t})^2\right)^{1/2}} \tag{7}$$

Unlike the test statistic in equation 6, our preferred test statistic in equation 7 uses the magnitude of the estimated average treatment effect across the post-treatment periods as the numerator. This better aligns with our interest in the likelihood of the average treatment effect being of the observed magnitude or greater. The fraction of test statistics in the permutation distribution greater than or equal to that of the treated unit's estimate is the p-value.

One challenge to inference in our setting is created by there being only up to 20 untreated districts in the analysis, limiting our ability to reliably assess the statistical significance of the

estimated effect for the treated unit.[51] We address this through the use of a bootstrap resampling method. Since each police district consists of smaller beats, we create resampled untreated districts by sampling beats with replacement from within each actual untreated district. These resampled districts are slightly perturbed versions of the originals and increase the number of untreated districts available for the inference procedure, while preserving the geographic clustering of beats by resampling within districts rather than across them, which may be important in environments in which statistical noise has a strong patterning by place and time.[52] For example, instead of an untreated district like the 2nd contributing a single value of the test statistic to the placebo distribution, we create 49 additional versions of the 2nd district by resampling with replacement from among its 15 beats and estimate treatment effects for each, contributing 50 values of the test statistic to the placebo distribution. We report p-values based on placebo distributions created both with and without this resampling procedure.

Finally, because we estimate the effects of SDSCs on multiple outcomes and units, we also report q-values that control the False Discovery Rate (FDR) based on the p-values obtained using the resampling procedure described above. Determining an estimate's significance using its p-value controls the false positive rate, or the share of truly null results we erroneously consider to be significant. With many tests, however, a large share of significant results will be truly null (false positives). In contrast, determining an estimate's significance using its q-value controls the FDR, or the share of significant estimates that are truly null results. We use the step-up procedure proposed by Benjamini et al. (2006) and the code written by Anderson (2008) to calculate these sharpened two-stage q-values within four outcome "families":[53] shootings; other reported crimes

[51] For example, suppose the test statistic for a treated unit is close to that for an untreated district. Under different realizations of the data, the ordering of these test statistics, and correspondingly the p-value, might change substantially.

[52] This is related to but slightly different from the approach employed by Robbins et al. (2017), who generate placebo areas using a permutation technique that groups together many comparison areas that are smaller (block-level) than their treatment area (neighborhood-level). As a result, their placebo areas are random assortments of small comparison areas that lack the structure of the treatment area, requiring them to standardize their effect estimates to guard against the resulting bias. In contrast, we avoid this issue by creating placebo districts using a resampling procedure, wherein we resample beats from within existing comparison districts with replacement, preserving the structure of those comparison districts in the process.

[53] The procedure proposed by Benjamini et al. (2006) controls the FDR when the test statistics are independent or

(violent felonies, Part 1 crimes, and all crimes); police activity measures (total arrests, arrests for guns, warrants, drugs and misdemeanors specifically, and traffic stops); and uses of force.[54]

4.2 Results for the 7th and 11th districts

We start with our short-run impact estimates for the 7th and 11th districts, the first two adopters of SDSCs in February 2017. Our preferred model specifications are chosen to maximize out-of-sample prediction accuracy by minimizing, for each outcome and treated unit, RMSPE in the $T_0 - T_0^{\text{placebo}}$ later pre-treatment months from a model estimated using data from the $T_0^{\text{placebo}} \geq \underline{T_0}^{\text{placebo}}$ earlier pre-treatment months, as described above and in Appendix B.1.

Because panel data models can sometimes be quite sensitive to even small estimation decisions, we also show how the results differ due to both changes in the model-specification selection process and across the model specifications themselves. Figure 7 shows how the preferred model specification for shooting victims in the 7th district changes under different choices of $\underline{T_0}^{\text{placebo}}$, the minimum number of months of pre-treatment data used to estimate the model in the backdating exercise. When using at least the first two years of pre-treatment data to estimate a model in the backdating exercise ($\underline{T_0}^{\text{placebo}} = 24$), the best-performing model uses the EN-M estimator, donor units drawn from Tiers 1-4 (excluding the 11th district), and donor units being districts. The pre-treatment RMSPE when estimated using the full dataset of this model (x-axis), and the resulting effect estimate (y-axis), are similar to those of other best-performing models chosen when $\underline{T_0}^{\text{placebo}} \leq 30$. However, when $\underline{T_0}^{\text{placebo}} \in 31, 32$, the best-performing model has both an unusually low pre-treatment RMSPE (suggesting it may be overfitting) and a much lower effect estimate. In addition to the sensitivity of the procedure for choosing a model specification, we also plot the distribution of all model specifications we estimate, with a vertical line indicating the one likeliest to maximize out-of-sample prediction accuracy by minimizing RMSPE in the $T_0 - T_0^{\text{placebo}}$ later

positively dependent.

[54] We separate shootings from other crimes to reflect that the SDSCs were designed and implemented specifically to reduce gun violence. We separate police activity measures from uses of force because many commanders might have as an explicit goal to increase activity, but presumably from society's perspective use of force is, all else equal, always desired to be lower.

pre-treatment months of the backdating exercise (Figure 8). Similar sensitivity analyses for each of the additional estimates presented below are shown in Appendix Figure 8.

Our results for the SDSCs' short-run impacts, reported in Table 4, are somewhat imprecisely estimated but consistent with the idea of large short-run reductions in gun violence and mixed impacts on enforcement harms like arrests. We find that the SDSCs reduced the rate of shooting victimization by 62% and 55% (column 4), respectively, in the 7th and 11th districts in February 2017. Stated differently, these estimates imply that there were 13 and 17 fewer shooting victims per 100,000 residents in the 7th and 11th districts (column 3), respectively, following their SDSCs' introduction in February 2017 than there would have been otherwise. Both estimates are significant at the 10% level before correcting for multiple testing (column 6), and fall just short of the 10% threshold after correcting (column 7). Figure 9 shows these estimates visually, as the difference in the post-treatment period between the solid black line (the actual rate of shooting victimization) and the dashed red line (the counterfactual rate of shooting victimization). The SDSCs' estimated effects on other measures of crime in February 2017—including violent felony offenses, Part I crimes, and all reported crimes—are generally imprecise for both districts.

While there are no statistically detectable changes in enforcement measures in the 7th district, in the 11th district we see signs of an increase in total arrests, equal to 26% ($p = 0.005$, and with our multiple-testing correction, $q = 0.031$). That increase in total arrests in the 11th district may have been driven by an increase in gun arrests, warrant arrests and misdemeanor arrests, all of which experienced large proportional changes (46%, 37%, and 26%, respectively) although only the estimated impact on misdemeanor arrests remains statistically significant after adjusting for multiple-testing ($q = 0.031$).[55]

Measuring *how* the SDSCs may have affected gun violence and enforcement harms is hampered by a lack of data on specific management practices like daily briefings and roll calls. Most police departments do not consistently track these sorts of internal process measures, perhaps in

[55] Note that, because the best-performing specification in the backdating exercise for total arrests and several other outcomes in the 11th district features a narrow donor pool consisting of only the other four Tier 1 districts (Appendix Table 4(a)), the lowest possible unadjusted p-value for these outcomes is 0.2.

part because of under-appreciation in the field about the importance of management practices. Unfortunately, we were unable to carry out original data collection to capture such measures partly due to budget and operational constraints. Instead, we examine three *intermediate outcomes* from administrative records that, while imperfect, provide some indirect evidence or insight into what the SDSCs were doing.

First, we measure rates of arrests initiated by officers monitoring CPD's POD cameras. Prior to the SDSCs, few officers outside of headquarters utilized the cameras, at least in part due to the cumbersome user interface required to do so. The SDSCs introduced new software that made monitoring cameras much easier. So rates of POD camera arrests offer a crude measure of how much commanders prioritized—and officers embraced—the use of new technologies. Appendix Figure 9 shows that the 7th and 11th districts saw large increases in rates of POD camera arrests immediately following their SDSCs' introduction in February 2017, while some other Tier 1 districts, notably the 9th and 10th, saw much smaller increases.[56] In addition to significant variation in POD camera arrest rates across districts, Appendix Figure 9 also shows large within-district variation in these rates over time. This suggests that districts may have been experimenting with this new technology, implying that there may be dynamic patterns to the SDSCs' treatment effects that are not captured in the short-run estimates presented here.

Second, we measure rates of officer self-reports of positive community interactions, or PCIs. Officers were encouraged by some commanders, particularly then-Commander Kenneth Johnson of the 7th district, to engage in and report positive—i.e., non-punitive, non-investigatory—interactions they had with members of the community, such as check-ins with local retailers or elderly wellness checks. The department began tracking PCIs as part of CompStat in 2017. Commander Johnson's emphasis on PCIs explains the large increase in the volume of PCIs in the 7th district following the SDSC's introduction, though officers from other SDSC districts followed suit in the second half of 2017 (Appendix Figure 10). We have no independent way to determine if officers were changing their behavior rather than simply changing their reporting. However, it is noteworthy that this

[56] Districts in Tiers 2 through 4 did not receive SDSCs—and the updated software to monitor cameras—until 2018 or later, and are therefore not expected to have increased rates of POD camera arrests in 2017.

increase was by far the largest in the 7th district, which is one of the districts with a reduction in gun violence that was notably above the Tier 1 average (discussed below) and which, unlike the 11th district, had no detectable changes in arrests or other enforcement measures.

Finally, we describe changes in spatial patterns of patrol activity in the 7th district following the introduction of its SDSC. Appendix Figure 11 shows two choropleths for the 7th district, each displaying the top 15% of cells by frequency of GPS pings during a period before the SDSC's introduction (Feb. - Dec. 2016) and after (Jan. - Dec. 2017). Prior to the SDSC, patrol activity throughout the 7th district was more dispersed, with more side streets represented among the top 15% of cells (Appendix Figure 11a). After the SDSC, patrol activity in the 7th district was more concentrated on major thoroughfares, particularly those running north-south (Halsted and Loomis streets together with Damen Avenue) (Appendix Figure 11b). Though only suggestive, this shift in patrol patterns may reflect the commander relying more on data provided by the SDSC to determine where resources are allocated.

4.3 Results for Tier 1 districts as a whole

The previous section reports short-run impact estimates of the SDSCs in February 2017 in the 7th and 11th districts. This section explores the SDSCs' impacts over longer time periods by combining data from the six Tier 1 districts, which received SDSCs in either February or March 2017, and constructing comparison groups using data from the Tier 2-4 districts, which did not begin receiving SDSCs until 2018. This approach allows us to estimate effects over a post-treatment window of up to 11 months (Feb. - Dec. 2017). Aggregating the Tier 1 districts rather than estimating effects separately for each prevents us from exploring heterogeneity in effect estimates, but it yields estimates of the SDSCs' average effects that are, on the whole, likely to be more reliable (see Appendix B.1).[57]

Results from this analysis are reported in Table 5, suggesting reductions in gun violence and

[57] For completeness, we report district-specific effect estimates for all Tier 1 districts for the post-treatment window through Dec. 2017 in Appendix B.2.

other crime through the first three months with no detectable changes in enforcement outcomes like arrests or police use of force. Shootings in the Tier 1 districts declined by 32% in February 2017, and by 25% and 21% through March and April, respectively (all $p < 0.05$). Figure 10 show these estimates visually at three and 11 months. Given Chicago's high degree of racial segregation and the demographic composition of the Tier 1 districts, this reduction in gun violence disproportionately benefits Black victims; if the 21% decline persisted through 2017, it would have reduced the city-wide Black-White disparity in shooting victimizations by 13%.[58] The SDSCs are estimated to have reduced reported violent felonies (by ~9%), Part I crimes (by ~7%), and all crimes (by ~7%) through three months, with most estimates remaining statistically significant after adjusting for multiple testing. Our estimates for impacts on overall and misdemeanor arrests in the first month are 11% and 8.7%, respectively; these are statistically significant considered on their own ($p < 0.05$) but not once we account for our multiple-testing corrections ($q > 0.1$).

However, by six months after the initial SDSCs were introduced, their estimated impacts on gun violence and other crime seem to attenuate, with smaller and less precise estimates, though we cannot rule out modest declines. Meanwhile, there are also signs that the composition of the arrests made in these Tier 1 districts changed. The estimated effect on total arrests by six and 11 months equaled 8.3% and 9.4%, but neither estimate approaches traditional statistical significance thresholds. But there is clear evidence that drug arrests increased, with statistically significant impacts of 72% and 58% at six and 11 months, as well as a statistically significant 19% increase in gun arrests at 11 months.

Taken together, the results through the first three months are consistent with a movement towards the PPF: outcomes that society wishes would decline, such as shooting victims and violent felonies, decline, while no other outcomes (at least that we can measure) that society cares about "get worse." This pattern begins to reverse by six months, with the effects on gun violence and other crimes appearing to attenuate. The results suggest that an intervention designed to

[58] The observed Black-White difference in rates of fatal plus non-fatal shooting victimizations in 2017 was 311 per 100,000. If we assume that victimization rates were 21% lower in the Tier 1 districts than they would have been absent the SDSCs, the counterfactual Black-White difference would have been 358 per 100,000.

improve implementation of CPD's existing strategies by introducing data-driven management practices in a police district—rather than increase resources or alter the policing strategy—may have affected police outputs as well, although raise the question of why the initial impacts did not persist.[59] One potential clue comes from some indication in the data that the SDSCs' impacts on gun violence may have been larger in the short-run in the 7th and 11th districts relative to the other Tier 1 districts, and more consistent and sustained in the 7th district than in other Tier 1 districts (Appendix B), suggesting that the SDSC impacts may have varied across the city. A candidate explanation for this variation, as we explore in the next section, is variability in the quality of the SDSCs' implementation.

5 Variability in the SDSCs' Implementation

To explore variability in the implementation of the SDSCs, we focus on officers' use of one component of the larger intervention: the introduction of a place-based predictive policing software named HunchLab.[60] HunchLab is a tool that uses historical CPD data on reported victimizations, shooting incidents, and gun arrests to predict the likelihood of different types of violent crime occurring in small areas ("boxes") within a police beat over the following 8-hour shift ("watch"). The software divides the entire city into a grid, such that the resolution of one box is approximately 300 x 300 meters.[61] Commanders were encouraged to instruct officers to spend an extra 10-15 minutes in the nearest box shown by HunchLab when not responding to calls for service.[62]

A noteworthy feature of HunchLab that we take advantage of here for research purposes is

[59] The evaluation of the SDSCs by Hollywood et al. (2019) finds that they reduced monthly counts of shootings across Tier 1 and 2 districts by 8.7%, and this reduction was particularly large in the 7th district. However, as mentioned earlier, their two-way fixed effects DD estimator does not address the measurement challenges posed by the SDSC districts—particularly the 7th and 11th—being extreme outliers, and is susceptible to significant bias in settings where treatment timing is staggered and treatment effects are heterogeneous (Goodman-Bacon, 2021).

[60] HunchLab was developed by Azavea and subsequently purchased by ShotSpotter, Inc. in 2018.

[61] An example of how this technology appeared to officers in Chicago's 12th district at the beginning of a watch is shown in Appendix Figure 12. Officers could also view the boxes on their newly issued phone using the HunchLab app, which included a timer for tracking the duration of time they spent in a box.

[62] Due to Chicago's high level of racial and economic segregation, and because HunchLab was used only to re-allocate officer time within rather than across police beats, there is little effect of HunchLab on the racial or economic composition of areas receiving more or fewer patrol resources.

that the boxes it shows at the start of each shift or watch are selected by a type of lottery. For each box at the start of each watch, HunchLab generates predictions for several types of violent crimes—homicides, shootings, robberies, assaults—that it aggregates and weights according to their estimated social cost, creating a single "weighted (predicted) risk" measure. Then, HunchLab chooses four boxes in each beat to show officers via a lottery, where boxes with a higher weighted risk relative to others in the same beat have a higher probability of being shown.[63]

We can use this randomization to help measure the fidelity of districts' implementation of the SDSCs. Specifically, we estimate the effects of HunchLab showing a box on how much time officers spend there in each district, and then we examine the degree to which variation in these estimated effects across districts are explained by variation in the estimated effects of officer time on gun violence. To do this, we use data from the full set of Tier 1 districts that adopted SDSCs, and specifically use data on 36,024 lotteries that took place between May 15, 2017 and October 31, 2017. Unlike in a conventional lottery, we cannot estimate effects simply by comparing outcomes in boxes that were and were not shown, because boxes that were shown have higher weighted risk, on average, than those not shown.[64] While we do not have access to the exact function that maps a box's relative weighted risk to its probability of being shown, we attempt to recover it from the data by constructing a moving average estimate of a box's probability of being shown across multiple lotteries. Let T represent the number of observed time periods in our dataset prior to the current period, $T+1$, and let I_{bt} for $t \in 1, \ldots, T$ be an indicator with a value of 1 if box b was shown in time t and 0 if it was not. Then our estimate of the probability of box b being shown to an officer at time $T+1$ is:

$$\hat{e}_{b,T+1} = \frac{\sum_{t=1}^{T} I_{bt}}{T} \tag{8}$$

Using this estimated probability, we can empirically confirm its positive relationship with a box's

[63] We believe the use of randomization is intended to increase compliance with HunchLab's recommendations by reducing the likelihood that officers are shown the same boxes watch after watch.

[64] To see this, Appendix Table 1 compares the characteristics of shown and not shown boxes in our dataset. Shown boxes in each district have significantly higher weighted risk than not shown boxes, with all pairwise differences being highly significant. If officers were already more likely to spend additional time in higher risk areas, then comparing time in shown and not shown boxes would overestimate the effect of HunchLab showing a box to officers.

relative weighted risk (Appendix Figure 13). We would like to compare shown and not shown boxes within the same lottery that have a similar probability of being shown. However, because most shown boxes lack a match to a not shown box with a similar probability of being shown within the same lottery (unique at the day, watch, and beat level), we instead search for matches across other watch-beat lotteries within the same week. In practice, our matching algorithm often matches a shown box during one day-watch to the same box during another day-watch in the same week; their predicted risk is similar, but whether the box is shown on one day or another is determined by the flip of a coin. The analysis sample of 34,884 shown boxes and an equal number of matched comparisons is approximately a quarter of the full experimental sample.[65]

Using the analysis sample, we first estimate the effects of a box being shown by HunchLab on officer time, a measure of SDSC implementation fidelity, both overall and by district using CPD vehicle GPS data.[66] We then use two-stage least squares (2SLS) to estimate the effects of officer time on gun violence, overall and by district, instrumenting for officer time using an indicator for whether a box was shown by HunchLab. To avoid potential reporting effects from greater police presence, we use a measure of gun violence unlikely to be affected: gunshots detected by ShotSpotter, an acoustic gunshot detection system installed in all SDSC districts.[67] We further recognize that, due to the small size of boxes, additional officer time in a given box may affect gun violence in surrounding boxes as well. Therefore, we define our gun violence outcome as the weighted sum of ShotSpotter alerts in a box and the immediately adjacent boxes, with the latter receiving half weight.[68] In all equations we control for the estimated probability of a box being shown; fixed effects for the interaction of week, watch, and beat; and fixed effects for day of the week. More formally, for box b within beat s in district d during week-watch w and day t, we

[65] The analysis sample is balanced on predicted risk (Appendix Table 2). To maximize the size of this analysis sample, we perform a grid search over several different matching parameters, including the caliper and the lowest and highest normalized weighted risks within a beat, subject to the criterion that boxes' weighted risks be balanced between those shown and not shown.

[66] These data exist as pings indicating the position and velocity of a vehicle at regular time intervals. We aggregate these pings to the box-day-watch level.

[67] Because ShotSpotter was installed at slightly different times across the Tier 1 districts, we restrict the 2SLS estimation to the subset of dates when ShotSpotter was active in a district.

[68] We also report results where the outcome is defined solely as ShotSpotter alerts in a box in Appendix Table 3.

estimate:

$$
\begin{aligned}
T_{bsdwt} &= \beta_0 + \beta_1 D_{bsdwt} + \beta_2 \hat{e}_{bsdwt} + \gamma_{sdw} + \sigma_t + \varepsilon_{bsdwt} \\
Y_{bsdwt} &= \alpha_0 + \alpha_1 T_{bsdwt} + \alpha_2 \hat{e}_{bsdwt} + \delta_{sdw} + \eta_t + \nu_{bsdwt}
\end{aligned}
\tag{9}
$$

where Y_{bsdwt} is a count of ShotSpotter alerts, T_{bsdwt} is officer time in minutes, $\hat{e}_{bsdwt}$ is the estimated probability of a box being shown, D_{bsdwt} is an indicator for whether the box was shown by HunchLab, γ_{sdw} and δ_{sdw} are week-watch-beat fixed effects, and σ_t and η_t are day-of-week fixed effects. The coefficients of interest, β_1 and α_1, measure the effect of a HunchLab box being shown on officer time and the effect of officer time on ShotSpotter alerts, respectively.

We see for starters that, across all of the Tier 1 districts (Table 6), having HunchLab show (flag) a box to officers increases officer time in that box by 0.2 minutes (column 2), or 2% of the mean among not shown boxes (column 1), a relatively precise estimate ($p = 0.013$). But this adoption of HunchLab to deploy officer time seems to have varied substantially across districts. In the 7$^{\text{th}}$ district—where we estimate both an unusually large reduction in shootings following the SDSC's introduction (Table 4) and whose commander's tenure during this period was characterized by both relatively lower shootings and lower uses of force (Figure 6 and Appendix Figure 3)—we find that a box being shown by HunchLab increases officer time in that box during a watch by 0.425 minutes ($p = 0.011$), or 3.7% of the mean time officers spent during a watch in matched boxes that were not shown, nearly twice the overall average among Tier 1 districts. We estimate an increase of similar magnitude in the 9$^{\text{th}}$ district as well ($p = 0.034$). An F-test rejects the null hypothesis that the district-specific estimated effects of a HunchLab box being shown on officer time are jointly zero ($p = 0.032$).

This variability in HunchLab adoption across Tier 1 districts raises the natural question of whether this might be due to variability across districts in the productivity of officer time in the HunchLab-selected locations. For starters, we can pool data from all districts and generate IV estimates of the average productivity of one additional minute of officer time. We find that this additional minute reduces ShotSpotter alerts by 0.0029 (column 6), or 22% of the mean alerts in not shown boxes (column 5), though we cannot reject the null that this effect is zero. We report a

95% confidence interval beneath the estimate using the tF procedure developed by Lee et al. (2021) to provide correct coverage for smaller values of first-stage F-statistics (column 3).[69] Focusing on the IV point estimate, it implies that the elasticity of ShotSpotter alerts to officer time equals -2.24. Comparisons to other studies are complicated by the fact that the most similar crime measure used in previous studies is murder, the vast majority of which are committed with guns in the U.S. (around 90% in Chicago, for example).[70] With that caveat in mind, our estimate falls towards the top of the range of previous estimated elasticities of murder with respect to police, which spans -0.7 to -2.7.[71] To the extent that our estimate differs from those in the previous literature, this may reflect either the difference in context across studies (Chicago versus national data) or the possibility that focusing resources on the highest-crime areas increases the marginal crime-prevention effect.[72]

A different way to see this is to plot deviations from each district's mean officer time and ShotSpotter alerts for shown and not shown boxes as in Figure 11, analogous to partial regression leverage plots in Figure 2 of Kling et al. (2007). This visual makes clear that the districts where officers spent more time in boxes saw the largest reductions in ShotSpotter alerts (a dose-response relationship).

We find no evidence that variation across districts in adoption of HunchLab can be explained by commanders having private information about heterogeneity in the effects of officer time on ShotSpotter alerts across districts. To test this hypothesis we compare district-level 'first stage'

[69] The procedure developed by Lee et al. (2021) involves constructing confidence intervals for IV estimates by scaling the second-stage standard error by an adjustment factor based on the first-stage F-statistic. Because they provide these adjustment factors only for F-statistics larger than 4, we cannot report tF-adjusted 95% confidence intervals for each district-specific IV estimate of officer time productivity.

[70] There is a separate question of how accurate ShotSpotter alerts are. External validations suffer from the problem of a lack of a "ground truth" benchmark for comparison. To the extent to which ShotSpotter alerts are a noisy proxy for the true prevalence of gun violence, if the measurement error is of the classical variety, then that would have the consequence of reducing the precision of our estimates for the effect of extra police presence on shootings.

[71] This comes from Table 5 in Chalfin and McCrary (2018); the low-end elasticity is from the measurement error-corrected estimates from Chalfin and McCrary (2018), which we find more credible than ordinary least squares estimates that do not control for measurement error in light of the findings in Chalfin and McCrary (2018), while the high-end is from Lin (2009), which uses state sales tax as an instrument.

[72] Mohler et al. (2015) find that a different predictive policing tool, PredPol, seems to predict crime 1.4 to 2.2 times as well as human crime analysts. It is possible that most patrol activities nationwide are not focused on high-violence hot spots at all, so the average crime rate in the patrol locations in our study could be far higher than in the national samples examined in past studies.

estimates of HunchLab's effect on officer time and 'second stage' estimates of the effect of officer time on ShotSpotter alerts in Figure 12. There is no clear systematic relationship between these two estimates. A line fit through these points has a slope of 0.007, with a 95% bootstrap confidence interval of $[-0.342, 0.305]$.[73]

A final candidate explanation for variability in HunchLab adoption across districts is the possibility that commanders might be trying to optimize a richer objective function than just reducing shootings, for example reducing racial disparities in arrests or other enforcement measures. Concluding commanders were making a "mistake" in ignoring the HunchLab recommendations in that case would be an example of what Kleinberg et al. (2018) call *omitted payoff bias*. Yet in practice this is unlikely to explain the pattern of results we see here partly because the Tier 1 districts turn out to be fairly homogeneous with respect to the racial and ethnic composition of local residents.[74]

In sum, commanders who do not encourage or demand that their officers spend more time in HunchLab boxes (or do so successfully) are not doing so because they are in districts where HunchLab is less effective at recommending boxes that could benefit from additional patrol time, or as a result of other objectives besides gun-violence prevention. These commanders (or their officers) are instead simply not following the playbook.

6 Conclusion

One goal of this work is to stimulate more research on an under-appreciated, under-studied explanation for variation in policing outcomes: management quality, or departments' ability to implement their stated objectives with available resources. The field currently knows little about how management contributes to police productivity, which management practices might have the largest impact, or how best to change them. Some research underway, such as by Canales et al. (2020) involving police departments in Mexico, is beginning to change that. The evidence presented

[73] From 1,000 iterations of resampling pairs of shown and not shown boxes with replacement from within each district.

[74] For example Englewood (the 7th district) is 94% Black and 1% White, while East and West Garfield Park (the 11th district) is 95% Black.

here further suggests the potential value of better understanding implementation fidelity as a lever for changing policing outcomes.

Our analysis suggests that the influence of management practices could be quite large in practice. For example, we present evidence of the impacts of a set of management changes in Chicago police districts with a first-year cost of around $2 million, including start-up costs related to new technologies designed to facilitate policy planning and implementation monitoring. By way of comparison, the total CPD budget of $1.7 billion, 90% of which goes toward personnel costs, averages out to about $77 million per district per year. One limitation of our analysis is that we cannot measure the impacts of changes in policing on the full range of outcomes society might care about, such as public perceptions of police legitimacy, since those are not consistently captured in any existing data source. With that caveat in mind, and recognizing the limits of our causal analyses, we estimate that these management changes implemented in the highest-violence 'Tier 1' police districts of Chicago's South and West sides may have reduced shootings in the first three months of adoption by up to 21%.

This analysis takes the objectives of current police departments as fixed, set by the democratic processes operating in each city (as imperfect as they may be) without a normative view as to what the objectives of policing *should* be. We recognize there is widespread and deeply felt disagreement about the goals of policing, and perhaps even some concern about the consequences of improving the management of police departments in situations where people disagree with a given department's goals. Put differently, some observers may view improved management in policing as a negative, rather than a positive, development. However, if society's goals for policing change, then poor implementation and management could become a key barrier to the successful adoption of those changes and a force for preserving the status quo. The value of understanding the role of management quality, and how to change it, transcends the specific normative views any individual person or jurisdiction has for policing.

The magnitude of the changes we document here, which occurred without the districts receiving an influx of officers or adopting a novel policing strategy, suggests that management differences

could explain a sizable part of the divergence in gun violence since the early 1990s between New York and Los Angeles on the one hand and Chicago on the other (Figure 1). Over this time period, Chicago had levels of police resources similar to or greater than those of its peers, and adopted a similar high-level policing strategy. But New York and Los Angeles are widely recognized to have been among the leaders in implementing data-driven management practices and other attempts to professionalize their departments, the basic, unglamorous work required of any organization to fulfill its mission. It may or may not be a coincidence that both departments shared a police commissioner over this period (William Bratton) who reports that one of his goals was to professionalize these departments (Bratton and Knobler, 2021),[75] or that both departments had less residual variation in homicides and police killings of civilians than Chicago (Figure 3). Closing even a quarter of the gap in homicide rates we've seen between Chicago and New York over the past three decades would have resulted in 2,500 fewer murder victims in Chicago, disproportionately from Black and brown communities of the city.

Management changes may also represent a relatively low-cost way to improve policing outcomes for budget-constrained cities across the country. This finding is not just of scientific interest but should be of policy interest as well, given the scale of problems facing policing in the U.S. and the limited resources with which to address them: a 30% increase in homicides from 2019 to 2020 that seemed to continue into 2021, disproportionately affecting communities of color; a rate of police killings of civilians that, depending on the measure used, either remained steady or increased in recent years, also disproportionately affecting communities of color; and $4 trillion in unfunded pension obligations faced by local governments that are unlikely to just go away.[76]

[75] See, e.g., https://som.yale.edu/sites/default/files/files/Case_Bratton_2nd_ed_Final_and_Complete.pdf and http://lapd-assets.lapdonline.org/assets/pdf/Harvard-LAPD%20Study.pdf.

[76] See, e.g., https://insight.kellogg.northwestern.edu/article/the_impending_pension_problem

References

Abadie, Alberto (2021) "Using synthetic controls: Feasibility, data requirements, and methodological aspects," *Journal of Economic Literature*, 59 (2), 391–425.

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller (2010) "Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program," *Journal of the American Statistical Association*, 105 (490), 493–505.

——— (2015) "Comparative politics and the synthetic control method," *American Journal of Political Science*, 59 (2), 495–510.

Abadie, Alberto and Javier Gardeazabal (2003) "The economic costs of conflict: A case study of the Basque Country," *American Economic Review*, 93 (1), 113–132.

Abrams, David S, Marianne Bertrand, and Sendhil Mullainathan (2012) "Do judges vary in their treatment of race?" *The Journal of Legal Studies*, 41 (2), 347–383.

Abrams, David S and Albert H Yoon (2007) "The luck of the draw: Using random case assignment to investigate attorney ability," *U. Chi. L. Rev.*, 74, 1145.

Anderson, Michael 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.

Arnold, David, Will S Dobbie, and Peter Hull (2020) "Measuring racial discrimination in bail decisions," Working Paper 26999, National Bureau of Economic Research.

Athey, Susan, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi (2018) "Matrix Completion Methods for Causal Panel Data Models," Working Paper 25132, National Bureau of Economic Research.

Bartelsman, Eric, John Haltiwanger, and Stefano Scarpetta (2013) "Cross-country differences in productivity: The role of allocation and selection," *American Economic Review*, 103 (1), 305–34.

Bell, Monica C (2021) "Next-Generation Policing Research: Three Propositions," *Journal of Economic Perspectives*, 35 (4), 29–48.

Ben-Michael, Eli, Avi Feller, and Jesse Rothstein (2021) "The augmented synthetic control method," *Journal of the American Statistical Association* (just-accepted), 1–34.

Benjamini, Yoav, Abba M Krieger, and Daniel Yekutieli (2006) "Adaptive linear step-up procedures that control the false discovery rate," *Biometrika*, 93 (3), 491–507.

Berry, Christopher R and Anthony Fowler (2021) "Leadership or luck? Randomization inference for leader effects in politics, business, and sports," *Science advances*, 7 (4).

Bloom, Nicholas, Erik Brynjolfsson, Lucia Foster, Ron S Jarmin, Megha Patnaik, Itay Saporta-Eksten, and John Van Reenen (2017) "What Drives Differences in Management?" Working Paper 23300, National Bureau of Economic Research, http://www.nber.org/papers/w23300.

Bloom, Nicholas, Benn Eifert, Aprajit Mahajan, David McKenzie, and John Roberts (2013) "Does management matter? Evidence from India," *The Quarterly Journal of Economics*, 128 (1), 1–51.

Bloom, Nicholas, Raffaella Sadun, and John Van Reenen (2016) "Management as a Technology?" Working Paper 22327, National Bureau of Economic Research.

Bloom, Nicholas and John Van Reenen (2007) "Measuring and explaining management practices across firms and countries," *The Quarterly Journal of Economics*, 122 (4), 1351–1408.

——— (2010) "Why do management practices differ across firms and countries?" *Journal of Economic Perspectives*, 24 (1), 203–24.

Branch, Gregory F, Eric A Hanushek, and Steven G Rivkin (2009) *Estimating principal effectiveness*: Urban Institute.

Brantingham, P Jeffrey, Matthew Valasik, and George O Mohler (2018) "Does predictive policing lead to biased arrests? Results from a randomized controlled trial," *Statistics and public policy*, 5 (1), 1–6.

Bratton, William and Peter Knobler (2021) *The Profession: A Memoir of Community, Race, and the Arc of Policing in America*: Penguin Publishing Group.

Canales, Rodrigo, Jessica Zarkin, and Cosmo Gabaglio (2020) "The Importance of Managerial Quality for Police Organizations."

Cassell, Paul G and Richard Fowles (2018) "What Caused the 2016 Chicago Homicide Spike: an Empirical Examination of the ACLU Effect and the Role of Stop and Frisks in Preventing Gun Violence," *U. Ill. L. Rev.*, 1581.

Chalfin, Aaron, Benjamin Hansen, Emily K Weisburst, Morgan C Williams et al. (2020) "Police Force Size and Civilian Race," Working Paper 28202, National Bureau of Economic Research.

Chalfin, Aaron and Justin McCrary (2018) "Are US cities underpoliced? Theory and evidence," *Review of Economics and Statistics*, 100 (1), 167–186.

Chandrasekher, Andrea Cann (2016) "The Effect of Police Slowdowns on Crime," *American Law and Economics Review*, 18 (2).

——— (2017) "Police Labor Unrest and Lengthy Contract Negotiations: Does Police Misconduct Increase with Time Spent Out of Contract?," https://ssrn.com/abstract=2470344.

Clark, Damon, Paco Martorell, and Jonah Rockoff (2009) "School Principals and School Performance. Working Paper 38.," *National Center for Analysis of longitudinal data in Education research*.

Devi, Tanaya and Roland G Fryer (2020) "Policing the Police: The Impact of "Pattern-or-Practice" Investigations on Crime," Working Paper 27324, National Bureau of Economic Research.

Di Tella, Rafael and Ernesto Schargrodsky (2004) "Do police reduce crime? Estimates using the allocation of police forces after a terrorist attack," *American Economic Review*, 94 (1), 115–133.

Doudchenko, Nikolay and Guido W. Imbens (2017) "Balancing, Regression, Difference-In-Differences and Synthetic Control Methods: A Synthesis."

Draca, Mirko, Stephen Machin, and Robert Witt (2011) "Panic on the streets of London: Police, crime, and the July 2005 terror attacks," *American Economic Review*, 101 (5), 2157–81.

Durlauf, Steven N and Daniel S Nagin (2011) "Imprisonment and crime: Can both be reduced?" *Criminology & Public Policy*, 10 (1), 13–54.

Easterly, William and Steven Pennings (2020) "Leader value added: Assessing the growth contribution of individual national leaders," Working Paper 27153, National Bureau of Economic Research.

Evans, William N and Emily G Owens (2007) "COPS and Crime," *Journal of Public Economics*, 91 (1-2), 181–201.

Ferguson, Andrew Guthrie (2016) "Policing predictive policing," *Wash. UL Rev.*, 94, 1109.

Fryer, Roland G (2020) "An Empirical Analysis of Racial Differences in Police Use of Force: A Response," *Journal of Political Economy*, 128 (10), 4003–4008.

Garicano, Luis and Paul Heaton (2010) "Information technology, organization, and productivity in the public sector: Evidence from police departments," *Journal of Labor Economics*, 28 (1), 167–201.

Gibbons, Robert S. and John Roberts eds. (2012) *The Handbook of Organizational Economics*: Princeton University Press.

Goncalves, Felipe and Steven Mello (2021) "A few bad apples? Racial bias in policing," *American Economic Review*, 111 (5), 1406–41.

Goodman-Bacon, Andrew (2021) "Difference-in-differences with variation in treatment timing," *Journal of Econometrics*.

Grissom, Jason A and Brendan Bartanen (2019) "Principal effectiveness and principal turnover," *Education Finance and Policy*, 14 (3), 355–382.

Hagedorn, John, Roberto Aspholm, Teresa Córdova, Andrew V. Papachristos, and Lance Williams (2019) "The Fracturing of Gangs and Violence in Chicago: A Research-Based Reorientation of Violence Prevention and Intervention Policy,"Technical report, https://greatcities.uic.edu/wp-content/uploads/2019/01/The_Fracturing_of_Gangs_and_Violence_in_Chicago.pdf.

Harcourt, Bernard E (2009) *Illusion of order: The false promise of broken windows policing*: Harvard University Press.

Harcourt, Bernard E and Jens Ludwig (2006) "Broken windows: New evidence from New York City and a five-city social experiment," *U. Chi. L. Rev.*, 73, 271.

——— (2007) "Reefer madness: Broken windows policing and misdemeanor marijuana arrests in New York City, 1989-2000," *Criminology & Pub. Pol'y*, 6, 165.

Hoekstra, Mark and CarlyWill Sloan (2020) "Does race matter for police use of force? Evidence from 911 calls," Working Paper 26774, National Bureau of Economic Research.

Hollywood, John S., Kenneth N. McKay, Dulani Woods, and Denis Agniel (2019) "Real-Time Crime Centers in Chicago,"Technical report, RAND Corporation, https://www.rand.org/

pubs/research_reports/RR3242.html.

Hsieh, Chang-Tai and Peter J Klenow (2009) "Misallocation and manufacturing TFP in China and India," *The Quarterly Journal of Economics*, 124 (4), 1403–1448.

Jackson, C Kirabo, Jonah E Rockoff, and Douglas O Staiger (2014) "Teacher effects and teacher-related policies," *Annu. Rev. Econ.*, 6 (1), 801–825.

Kaplan, Jacob (2020) "Jacob Kaplan's Concatenated Files: Uniform Crime Reporting Program Data: Offenses Known and Clearances by Arrest, 1960-2019.,"Technical report, Inter-university Consortium for Political and Social Research, Ann Arbor, MI, https://doi.org/10.3886/E100707V16.

Kapustin, Max, Jens Ludwig, Marc Punkay, Kimberley Smith, Lauren Speigel, and David Welgus (2017) "Gun violence in Chicago, 2016," *University of Chicago Crime Lab*.

Kaul, Ashok, Stefan Klößner, Gregor Pfeifer, and Manuel Schieler (2018) "Synthetic control methods: Never use all pre-intervention outcomes together with covariates."

King, Gary and Langche Zeng (2006) "The dangers of extreme counterfactuals," *Political Analysis*, 14 (2), 131–159, 10.1093/pan/mpj004.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan (2018) "Human decisions and machine predictions," *The Quarterly Journal of Economics*, 133 (1), 237–293.

Klick, Jonathan and Alexander Tabarrok (2005) "Using terror alert levels to estimate the effect of police on crime," *The Journal of Law and Economics*, 48 (1), 267–279.

Kling, Jeffrey R, Jeffrey B Liebman, and Lawrence F Katz (2007) "Experimental Analysis of Neighborhood Effects," *Econometrica*, 75 (1), 83–119, https://doi.org/10.1111/j.1468-0262.2007.00733.x.

Lee, David S, Justin McCrary, Marcelo J Moreira, and Jack R Porter (2021) "Valid t-ratio Inference for IV," Working Paper 29124, National Bureau of Economic Research.

Levitt, Steven D (1997) "Using electoral cycles in police hiring to estimate the effects of police on crime," *American Economic Review*, 87 (3), 270–290.

——— (2002) "Using electoral cycles in police hiring to estimate the effects of police on crime: Reply," *American Economic Review*, 92 (4), 1244–1250.

Lin, Ming-Jen (2009) "More police, less crime: Evidence from US state data," *International Review of Law and Economics*, 29 (2), 73–80.

Lynn, Laurence E (1987) "Public management: What do we know? what should we know? and how will we know it?" *Journal of Policy Analysis and Management*, 7 (1), 178–187.

——— (2011) "Public management," *Journal of Policy Analysis and Management*.

MacDonald, John, Jeffrey Fagan, and Amanda Geller (2016) "The effects of local police surges on crime and arrests in New York City," *PLoS one*, 11 (6), e0157223.

Machin, Stephen and Olivier Marie (2011) "Crime and police resources: The street crime initiative," *Journal of the European Economic Association*, 9 (4), 678–701.

Mas, Alexandre (2006) "Pay, Reference Points, and Police Performance," *The Quarterly Journal of Economics*, 121 (3), https://academic.oup.com/qje/article-abstract/121/3/783/1917873.

Meares, Tracey (2008) "The legitimacy of police among young African-American men," *Marq. L. Rev.*, 92, 651.

Meares, Tracey L, Tom R Tyler, and Jacob Gardener (2015) "Lawful or fair-how cops and laypeople perceive good policing," *Journal of Criminal Law & Criminology*, 105, 297.

Mello, Steven (2019) "More COPS, less crime," *Journal of Public Economics*, 172, 174–200.

Mohler, George O, Martin B Short, Sean Malinowski, Mark Johnson, George E Tita, Andrea L Bertozzi, and P Jeffrey Brantingham (2015) "Randomized controlled field trials of predictive policing," *Journal of the American Statistical Association*, 110 (512), 1399–1411.

Mummolo, Jonathan (2018) "Modern police tactics, police-citizen interactions, and the prospects for reform," *Journal of Politics*, 80 (1), 1–15, 10.1086/694393.

National Academies of Sciences, Engineering, and Medicine (2018) *Proactive policing: Effects on crime and communities*: National Academies Press.

National Research Council (2004) *Fairness and effectiveness in policing: The evidence*: National Academies Press.

Olken, Benjamin A (2015) "Promises and perils of pre-analysis plans," *Journal of Economic Perspectives*, 29 (3), 61–80.

Owens, Emily G (2013) "COPS and Cuffs," *Lessons from the Economics of Crime: What Reduces Offending?*, 17.

Owens, Emily G, David Weisburd, Karen L Amendola, and Geoffrey P Alpert (2018) "Can you build a better cop? Experimental evidence on supervision, training, and policing in the community," *Criminology & Public Policy*, 17 (1), 41–87.

Robbins, Michael W, Jessica Saunders, and Beau Kilmer (2017) "A framework for synthetic control methods with high-dimensional, micro-level data: evaluating a neighborhood-specific crime intervention," *Journal of the American Statistical Association*, 112 (517), 109–126.

Rosenfeld, Richard and Robert Fornango (2017) "The relationship between crime and stop, question, and frisk rates in New York City neighborhoods," *Justice Quarterly*, 34 (6), 931–951.

Rubin, Donald B (1974) "Estimating causal effects of treatments in randomized and nonrandomized studies.," *Journal of Educational Psychology*, 66 (5), 688.

Shapiro, Aaron (2017) "Reform predictive policing," *Nature news*, 541 (7638), 458.

Skogan, Wesley G. and Susan M. Hartnett (1999) *Community policing, Chicago style*: Oxford University Press.

Skogan, Wesley G., Lynn Steiner, Jill DuBois, Erik K. Gudell, and Aimee Fagan (2002) "Taking

Stock: Community Policing in Chicago,"Technical report, National Institute of Justice.

Stone, Christopher, Todd S Foglesong, and Christine M Cole (2009) *Policing Los Angeles under a consent degree: The dynamics of change at the LAPD*: Program in Criminal Justice Policy and Management, Harvard Kennedy School.

Sullivan, Christopher M. and Zachary P. O'Keeffe (2017) "Evidence that curtailing proactive policing can reduce major crime," *Nature Human Behaviour*, 1 (10), 730–737, 10.1038/s41562-017-0211-5.

Syverson, Chad (2011) "What determines productivity?" *Journal of Economic literature*, 49 (2), 326–65.

Tyler, Tom R (2003) "Procedural justice, legitimacy, and the effective rule of law," *Crime and justice*, 30, 283–357.

Weisburd, David, Alese Wooditch, Sarit Weisburd, and Sue-Ming Yang (2016) "Do stop, question, and frisk practices deter crime? Evidence at microunits of space and time," *Criminology & public policy*, 15 (1), 31–56.

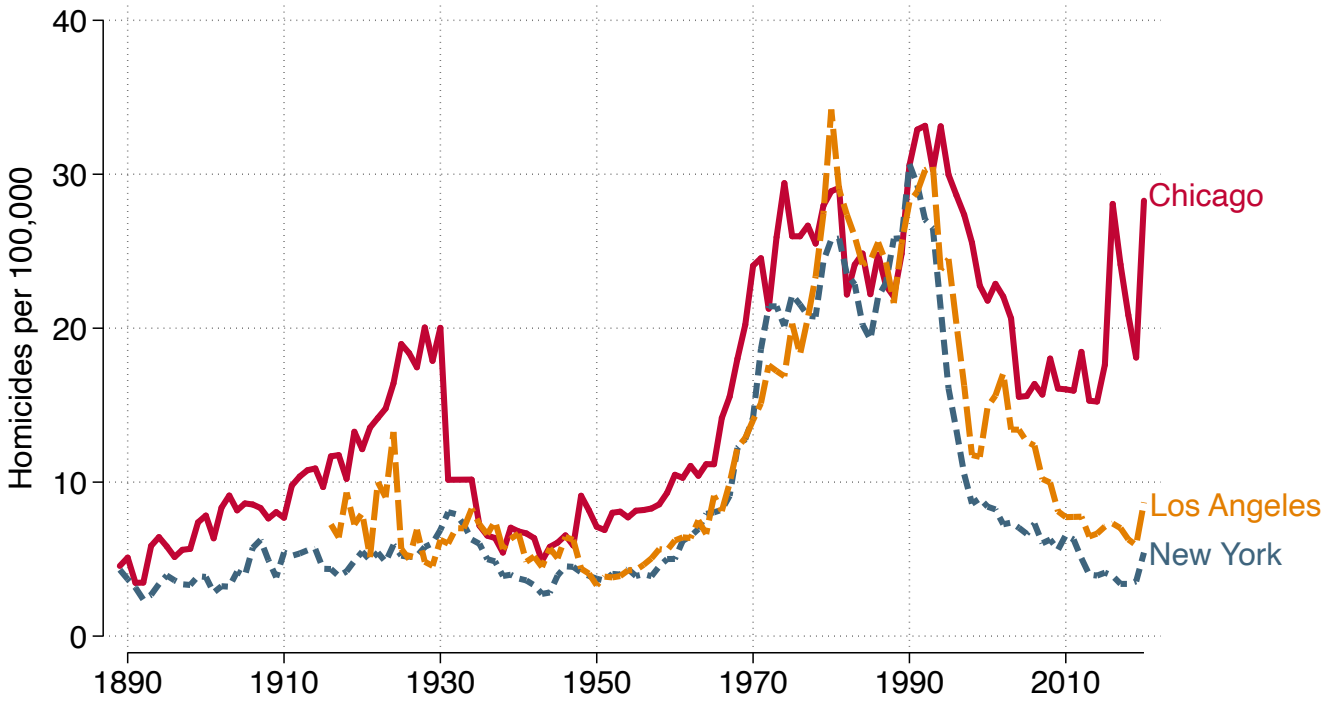
Weisburg, David, Roasann Greenspan, Stephen Mastrofski, James J Willis, and Police Foundation (2008) "Compstat and organizational change: A national assessment."

Wilson, James Q (1989) "Bureaucracy: What Government Agencies Do and Why They Do It."

Wolfers, Justin (2002) "Are Voters Rational? Evidence from Gubernatorial Elections."

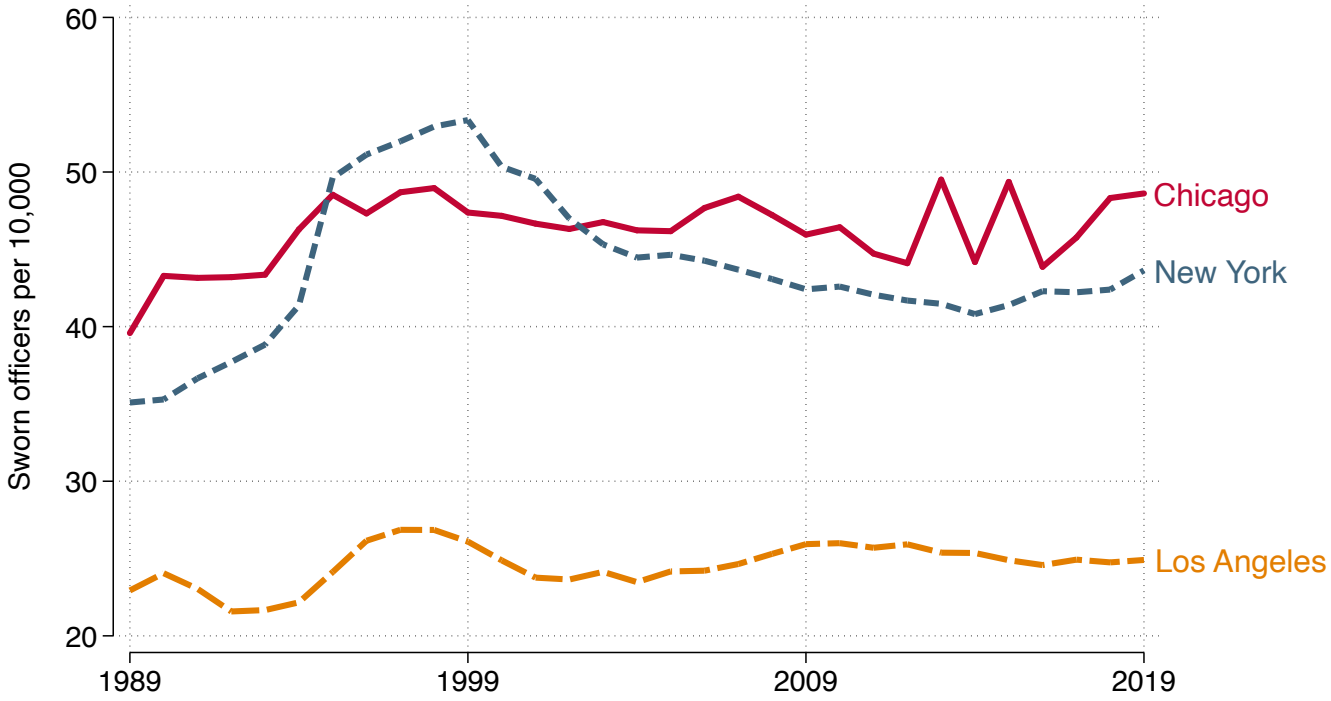
Zimring, Franklin E (2011) *The city that became safe: New York's lessons for urban crime and its control*: Oxford University Press.

Figure 1: Homicide rates in New York City, Los Angeles, and Chicago, 1889–2020



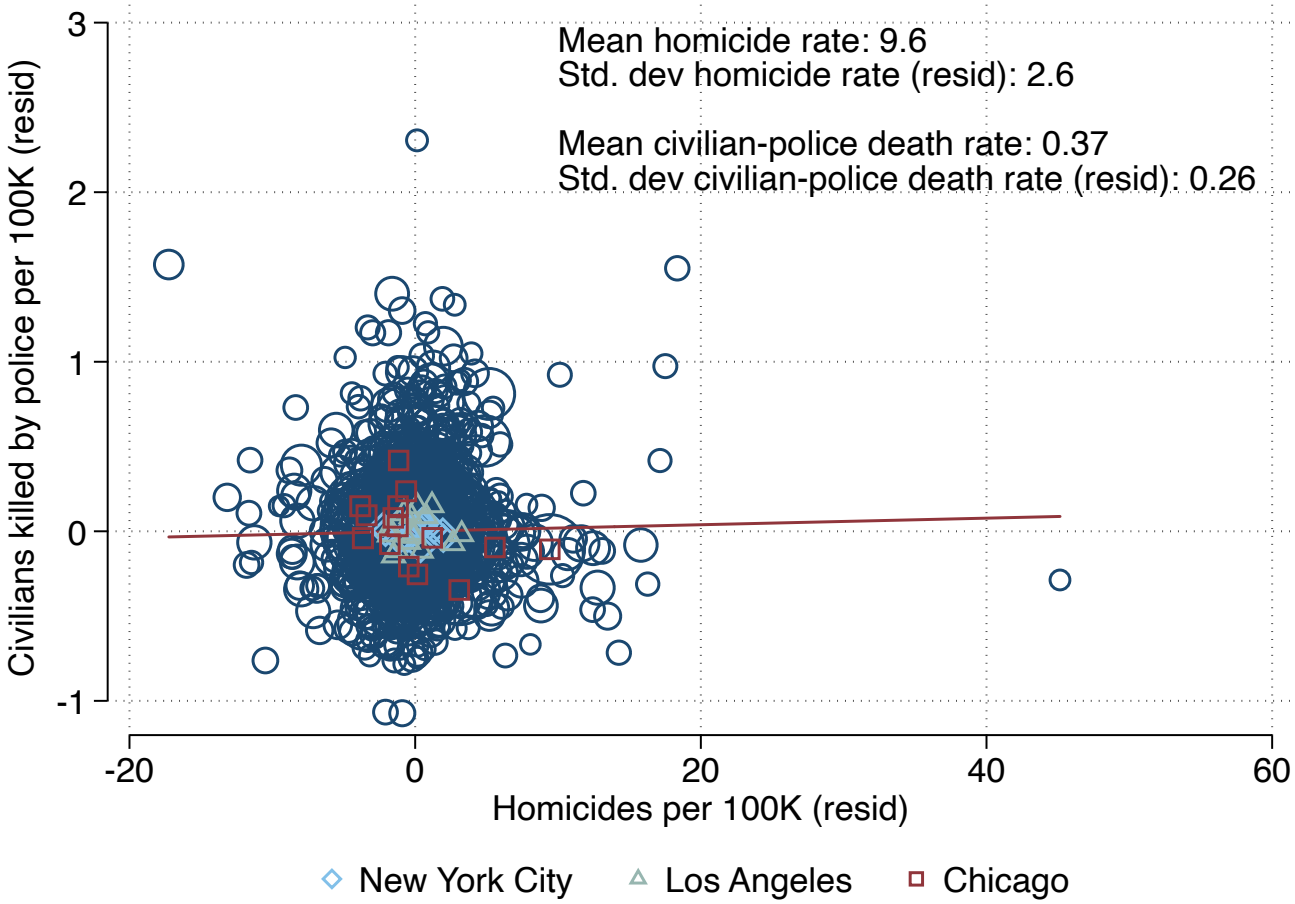
Note: Chicago homicide data for 1889 through 1930 from the Chicago Historical Homicide Project at Northwestern University. Chicago homicide data for 1930 through 1959 from the FBI's Uniform Crime Reports (ICPSR 3666). Los Angeles homicide data for 1916 through 1959 from the Historical Violence Database at the Criminal Justice Research Center, the Ohio State University. New York City homicide data for 1890 through 1959 from the National Institute of Justice (ICPSR 3226). Homicide data for 1960 through 2019 from the FBI's Uniform Crime Reports (Open ICPSR). Homicide data for 2020 from the police departments of Chicago, New York City, and Los Angeles.

Figure 2: Police officers in New York City, Los Angeles, and Chicago, 1989–2019



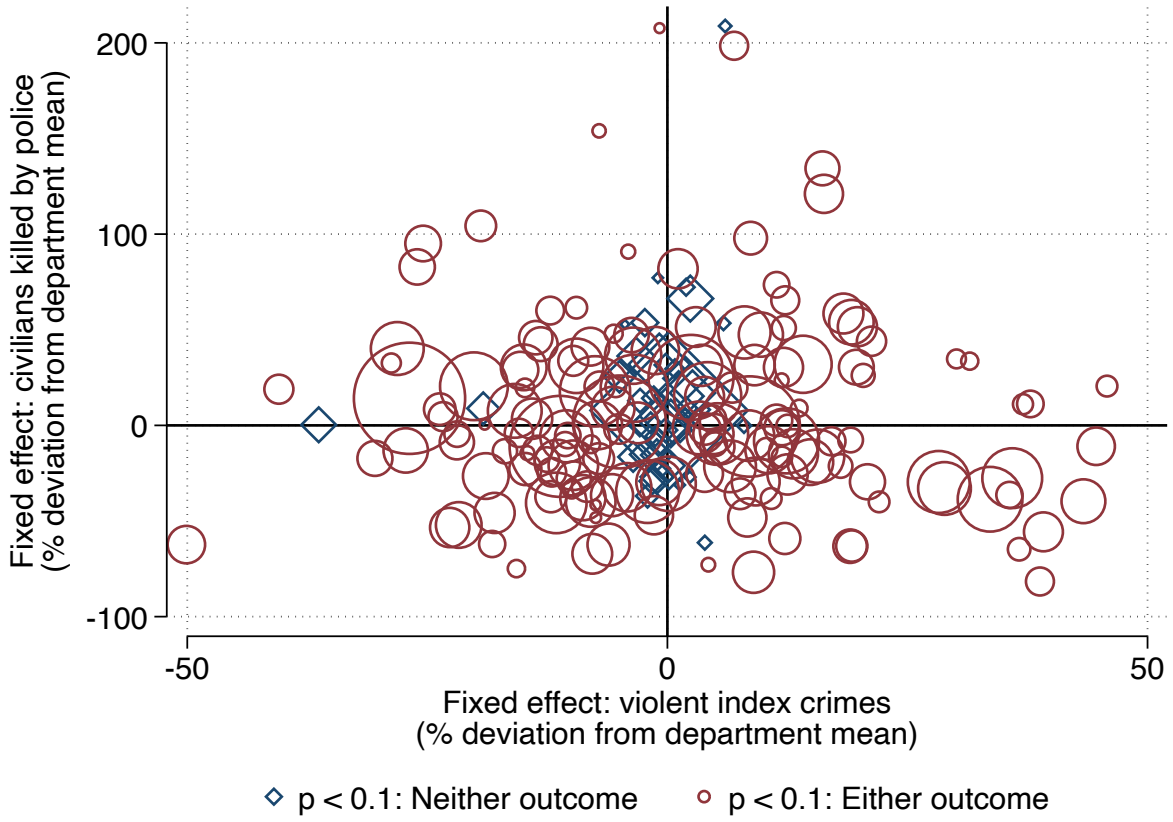
Note: Data from UCR LEOKA and NYPD Office of Management Analysis and Planning (OMAP). NYPD sworn staffing levels from 1990-2009 are based on OMAP data made available by Franklin Zimring (https://global.oup.com/us/companion.websites/9780199844425/). For discussion of errors in NYPD's sworn staffing levels in UCR data, see Chalfin and McCrary (2018).

Figure 3: Cross-city variation in police outputs: 50 largest cities, 2010–2019



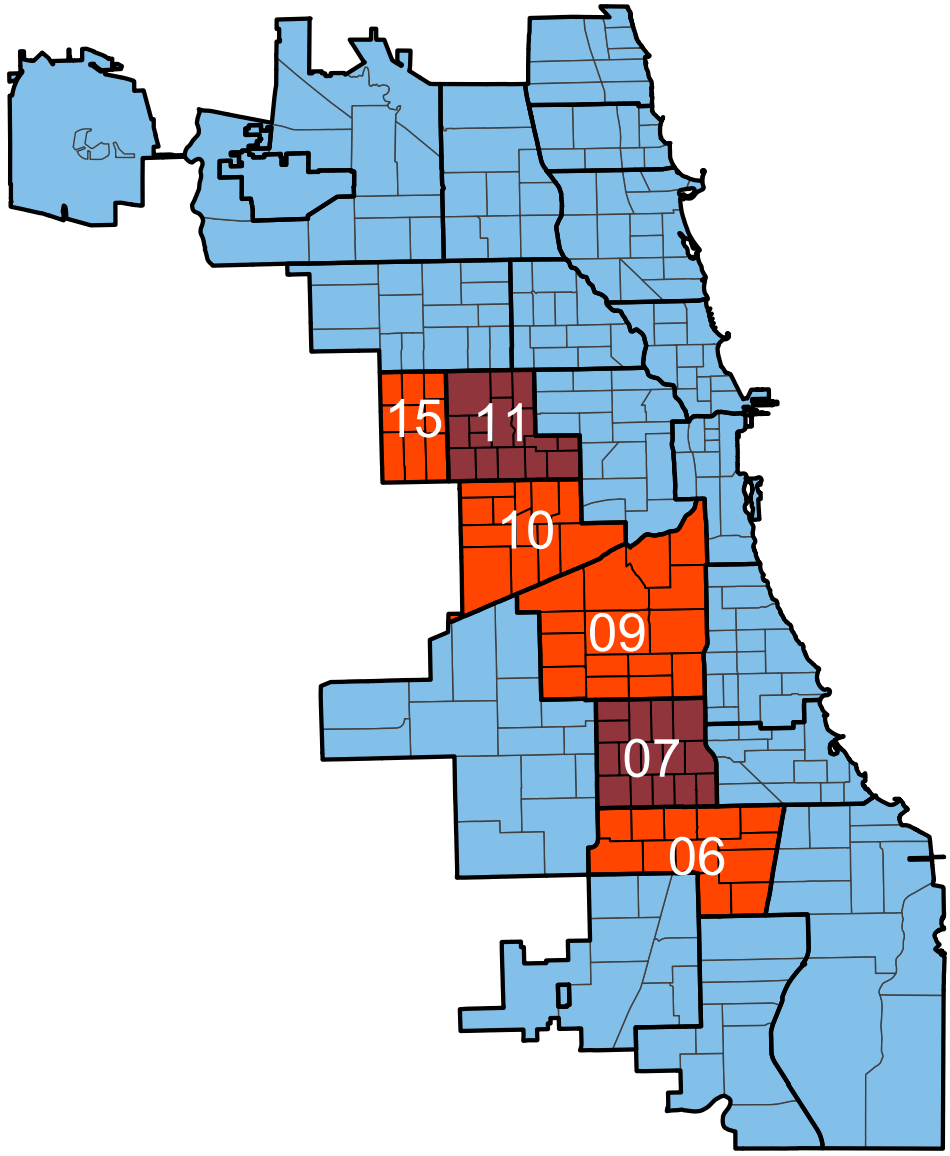
Note: Data from UCR and Fatal Encounters. Departments serving the 50 largest jurisdictions based on median population in 2010-2019. Each point is a pair of population-weighted residuals from estimation of equation 1. On the x-axis are residuals of a department's rate of homicides per 100,000 from the UCR, and on the y-axis are residuals of a department's civilians killed by police per 100,000 from Fatal Encounters. A population-weighted best-fit line through these points is shown. Text reports the population-weighted mean of each (unresidualized) outcome and standard deviation of the residuals. Finally, the points for New York City, Los Angeles, and Chicago are labeled separately.

Figure 4: Within-city, across-chief variation: tenure fixed effects



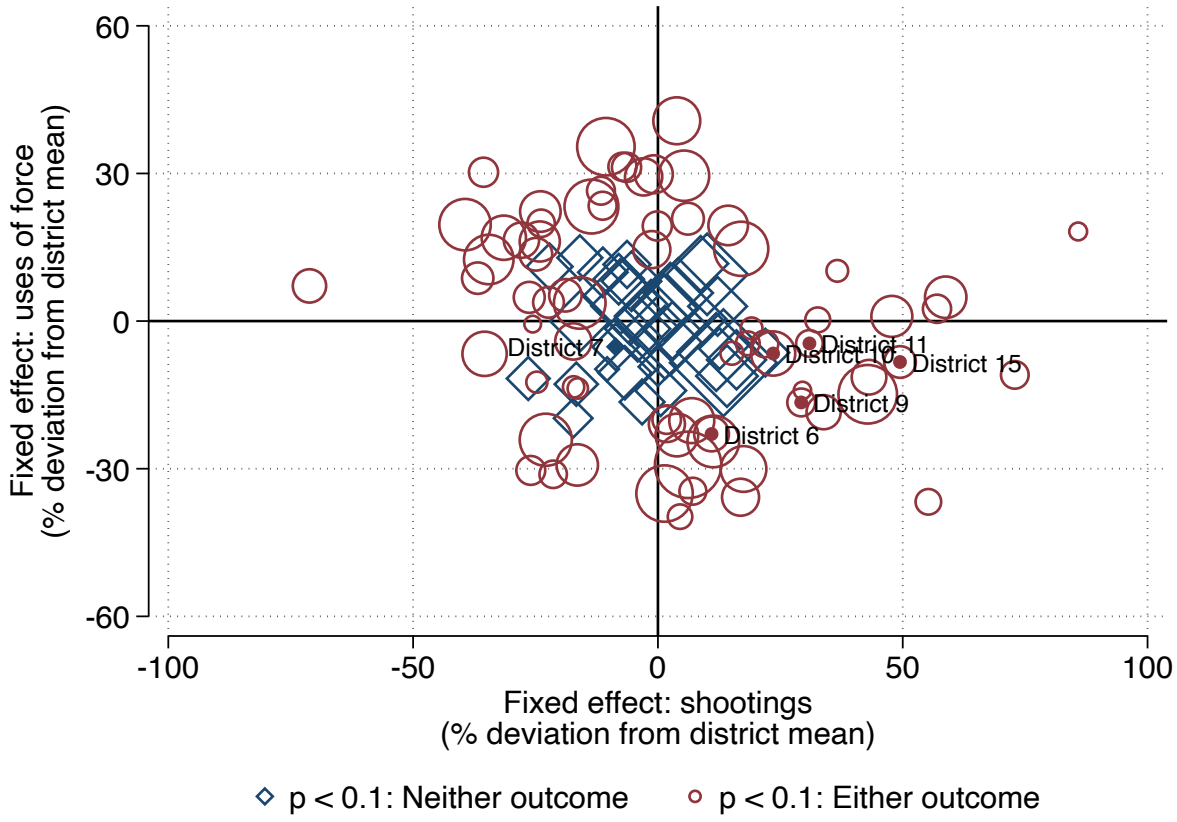
Note: Data from UCR and Fatal Encounters. Departments serving the 50 largest jurisdictions based on median population in 2010-2019. Each point is a pair of population-weighted police chief fixed effects estimates from equation 3. The sizes of the points reflect weighting for both jurisdiction population and tenure length. Fixed effects are not estimated for tenures of less than 6 months or chiefs who served in an interim capacity. Points where either fixed effect estimate is above (below) the 99th (1st) percentile are not plotted. On the x-axis is the average percent deviation in violent index crime rates from the department mean during a chief's tenure. On the y-axis is the average percent deviation in civilian-police death rates from the department mean during a chief's tenure. Hollow diamonds are tenures where neither fixed effect estimate has $p < 0.1$. Hollow circles are tenures where one or both fixed effect estimates has $p < 0.1$.

Figure 5: Chicago police districts and beats



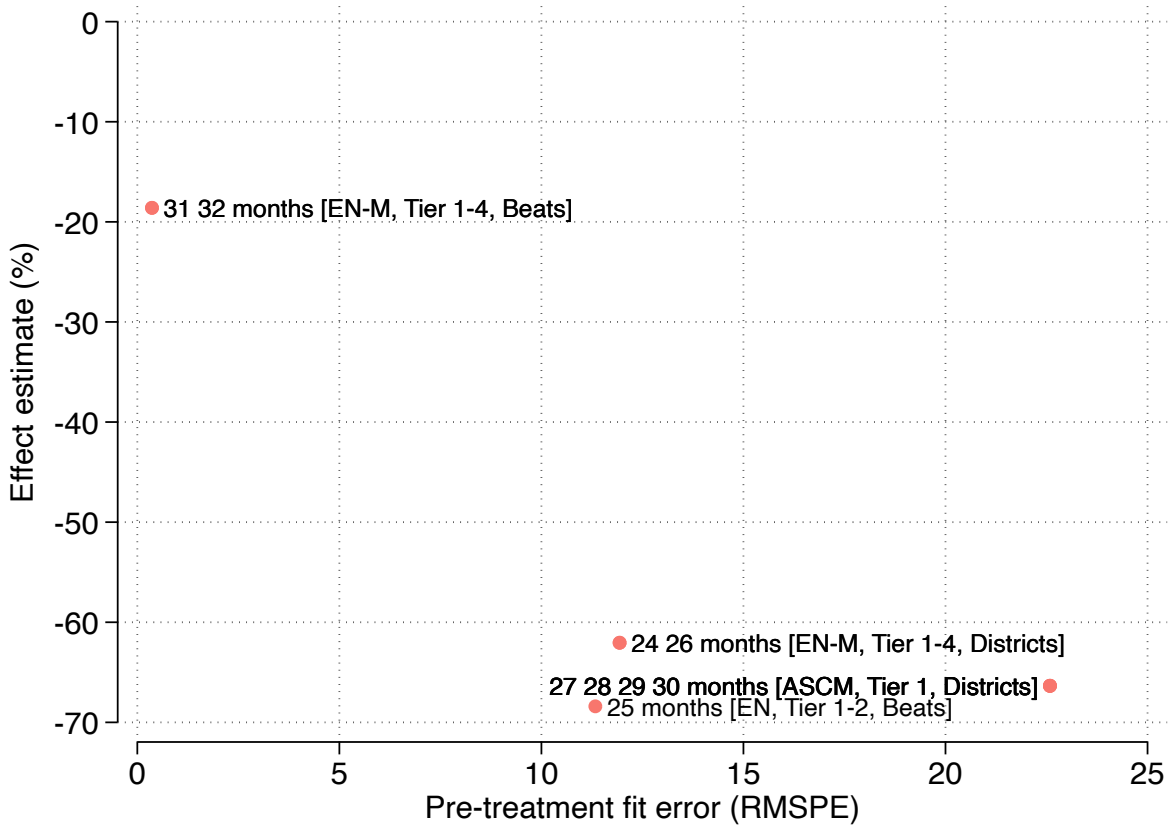
Note: Boundaries of Chicago's 22 police districts (bold) and their beats. The six Tier 1 districts are labeled, and the first two to receive SDSCs—the 7th and 11th—are shaded dark red.

Figure 6: Within-district, across-commander variation: tenure fixed effects



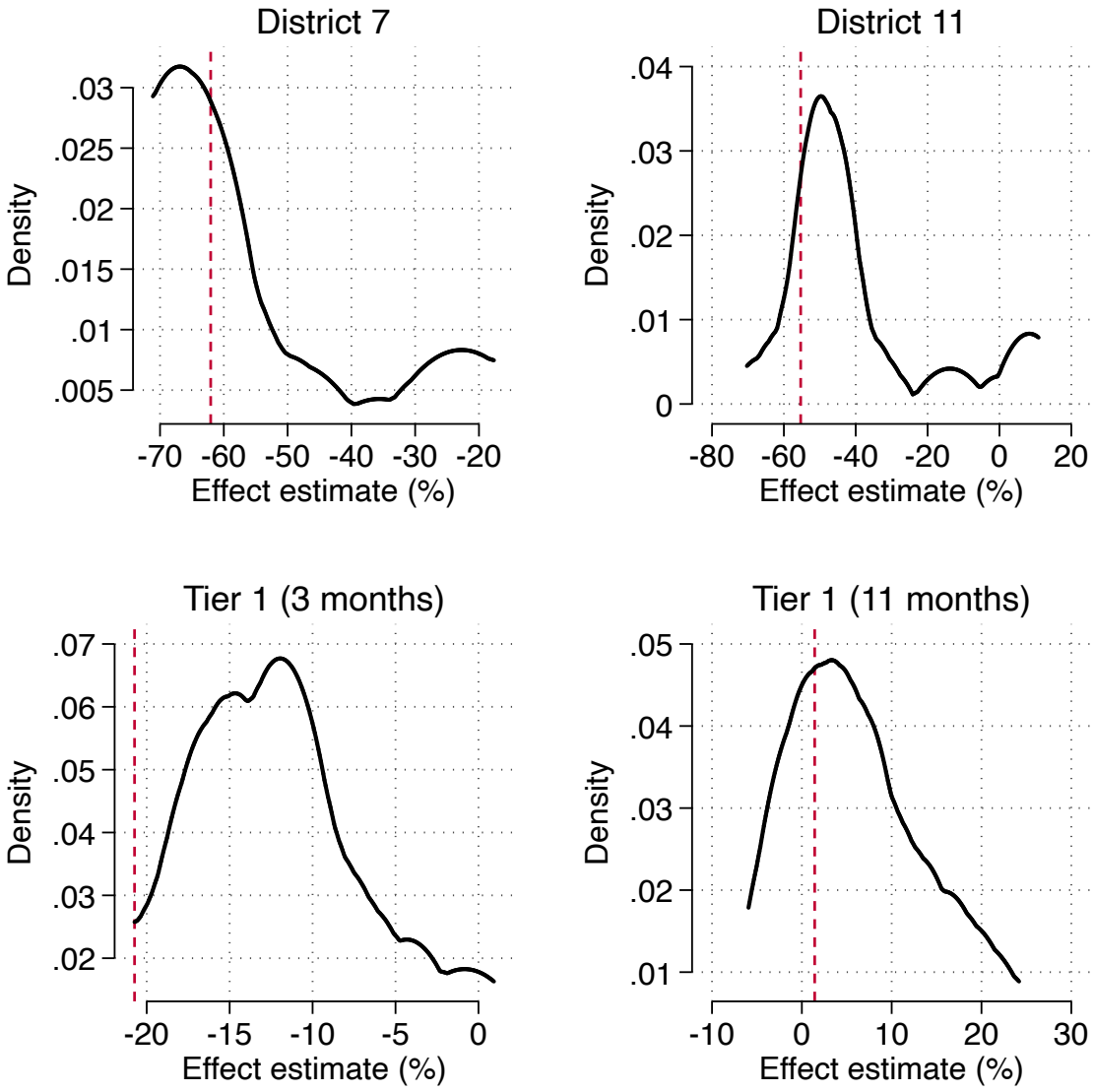
Note: Data from the Chicago Police Department. Each point is a pair of population-weighted district commander fixed effects estimates from equation 3. The sizes of the points reflect weighting for both district population and tenure length. Fixed effects are not estimated for tenures of less than 6 months or commanders who served in an interim capacity. Points where either fixed effect estimate is above (below) the 99^{th} (1^{st}) percentile are not plotted. On the x-axis is the average percent deviation in shooting rates from the district mean during a commander's tenure. On the y-axis is the average percent deviation in police use of force rates from the district mean during a commander's tenure. Hollow diamonds are tenures where neither fixed effect estimate has $p < 0.1$. Hollow circles are tenures where one or both fixed effect estimates has $p < 0.1$. The six labeled solid points represent the tenures of the commanders of the six Tier 1 districts during most or all of 2017.

Figure 7: Sensitivity of panel data model selection procedure: shooting victims in the 7th district



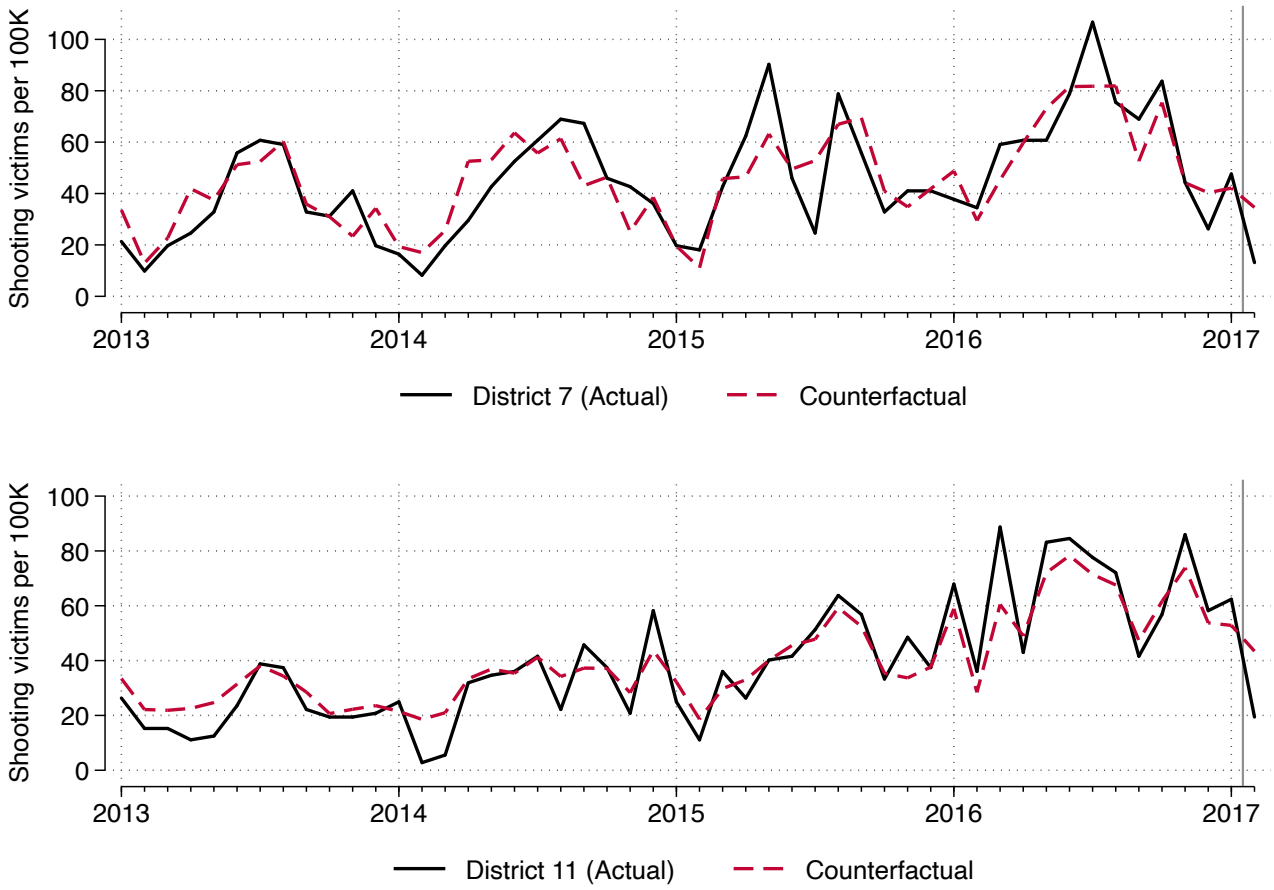
Note: Data from the Chicago Police Department. Each point represents a different model specification chosen to maximize out-of-sample prediction accuracy for estimating the impact of the SDSC in the 7th district on shooting victims in February 2017, depending on the value of $\underline{T_0}^{\text{placebo}}$, the minimum number of months of pre-treatment data used in the backdating exercise. There are $T_0 = 49$ months of pre-treatment data available, from January 2013 through January 2017. On the x-axis is the pre-treatment RMSPE of this model when estimated using the full 49 months of pre-treatment data. On the y-axis is the resulting point estimate of the SDSC's effect on shooting victims in February 2017. See section 4.2 and Appendix B.1 for additional details.

Figure 8: Distribution of effect estimates across model specifications: shooting victims



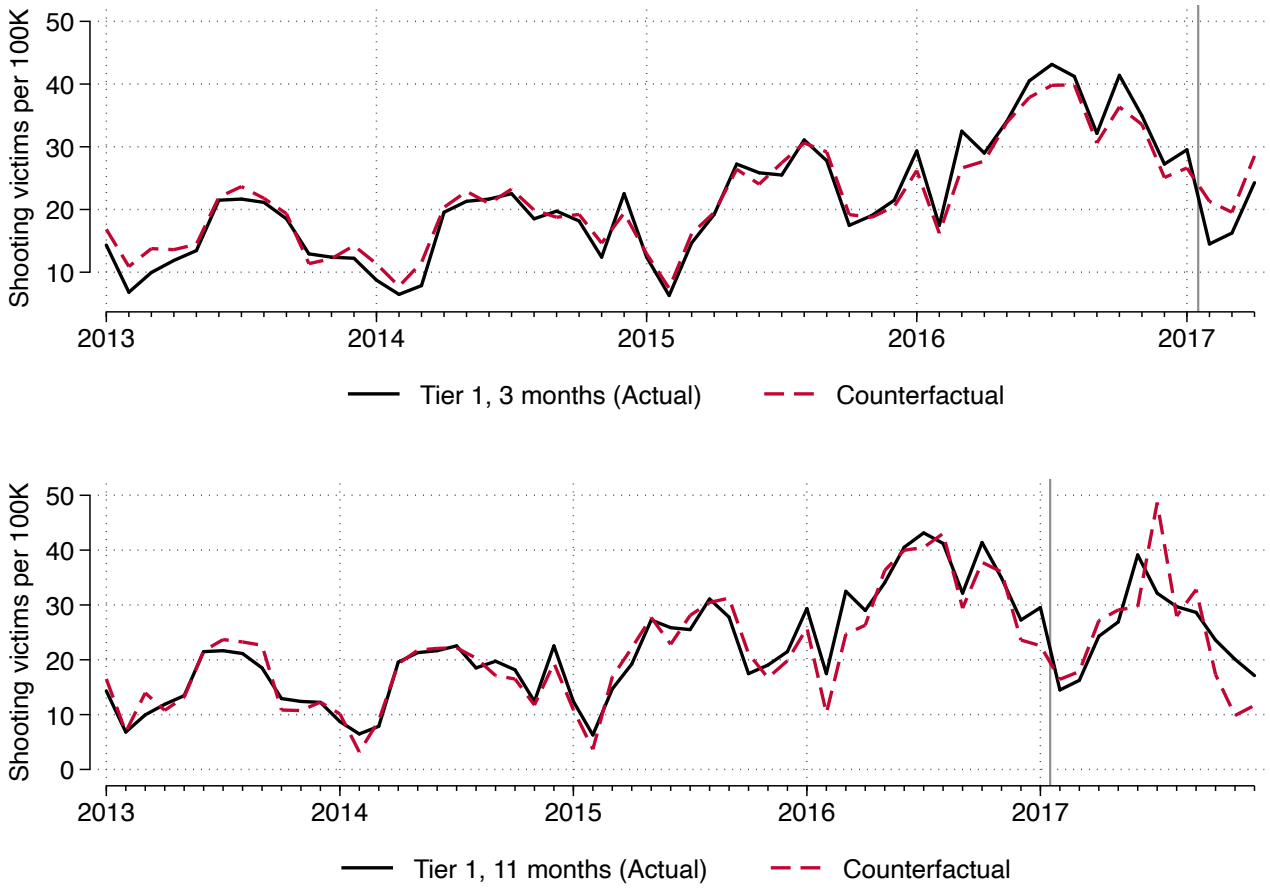
Note: Data from the Chicago Police Department. Each panel reports a distribution of estimated effects on shooting victims from different model specifications. The top two panels show distributions of estimated effects for February 2017 in the 7^{th} and 11^{th} districts, respectively (section 4.2). The bottom two panels show distributions of estimated effects for Feb. - Apr. 2017 and Feb. - Dec. 2017, respectively, for Tier 1 districts as a whole (section 4.3). Vertical lines correspond to the model specification that maximizes out-of-sample prediction accuracy in the backdating exercise.

Figure 9: Actual vs. counterfactual outcomes: 7th and 11th districts, short-run, shooting victims



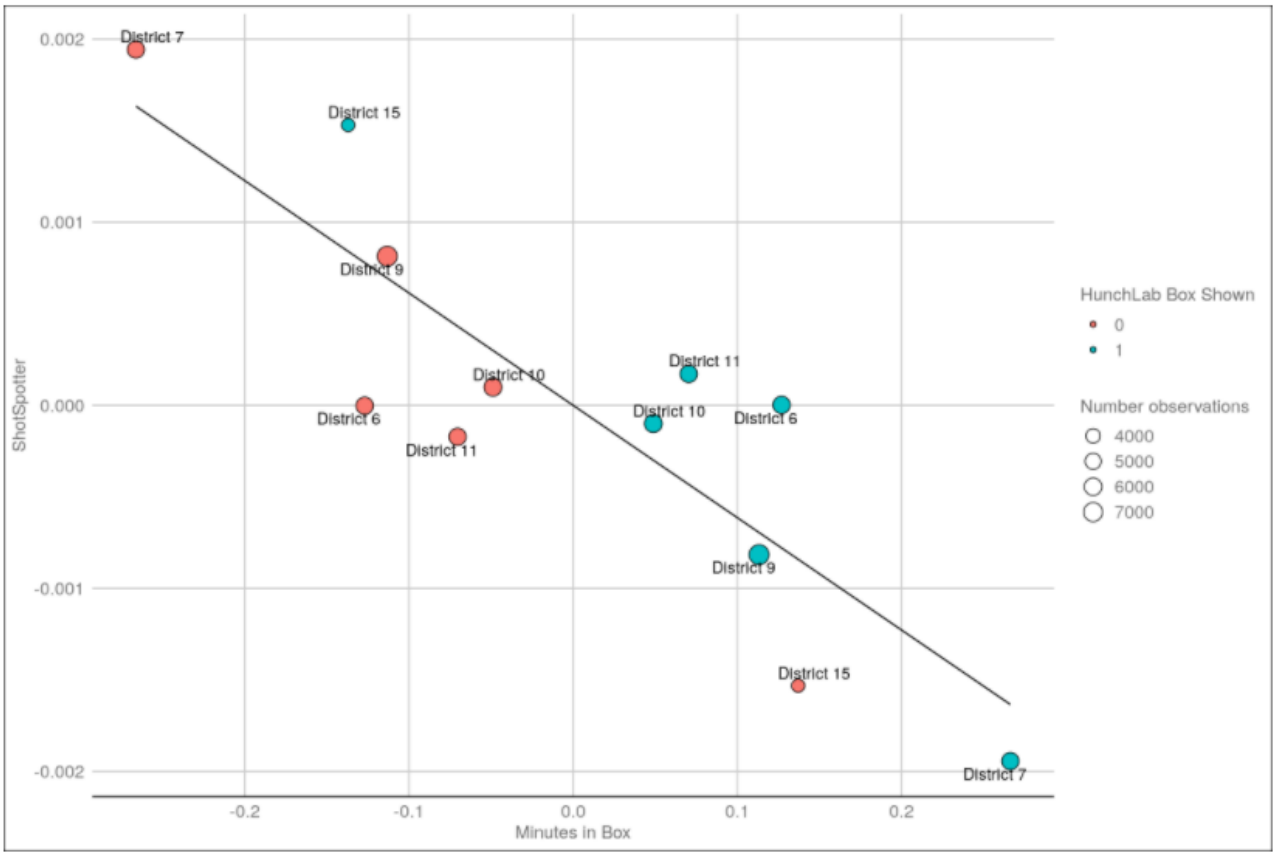
Note: Data from the Chicago Police Department. Panels show the actual number of shooting victims per 100,000, as well as the counterfactual estimate from the best-performing model specification, for the 7th district (top) and 11th district (bottom). Vertical lines indicate the launch of the each district's SDSC in February 2017.

Figure 10: Actual vs. counterfactual outcomes: Tier 1, shooting victims



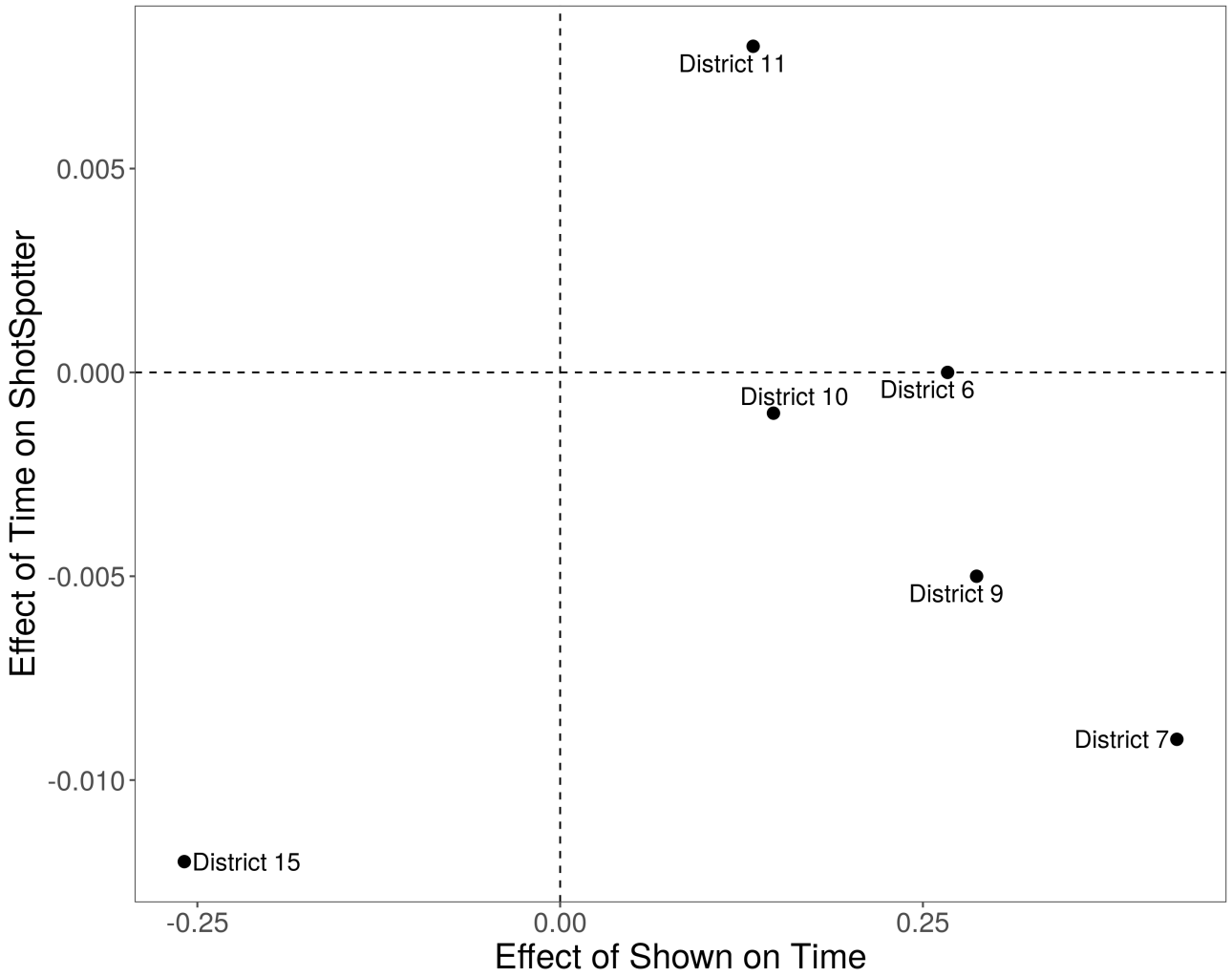
Note: Data from the Chicago Police Department. Panels show the actual number of shooting victims per 100,000, as well as the counterfactual estimate from the best-performing model specification, for Tier 1 districts combined, with an outcome window of three months (top) and 11 months (bottom). Vertical lines indicate the launch of the earliest SDSCs in February 2017.

Figure 11: Dose-response relationship between officer time and ShotSpotter alerts in HunchLab boxes



Note: Data from the Chicago Police Department. Each point represents the average deviation from the mean of officer time in a box (x-axis) and ShotSpotter alerts in a box (y-axis), by district, separately for boxes that were shown (blue) and boxes that were not shown (red) in the matched analysis sample. The size of each point is proportional to the number of observations it contains, and the weight each point receives in a two-stage least squares regression of ShotSpotter alerts on officer time in a box and district fixed effects, instrumenting for officer time with interactions of district fixed effects and whether a HunchLab box was shown. This is similar to the approach used to generate the partial regression leverage plots in Figure 2 of Kling et al. (2007).

Figure 12: HunchLab first- vs. second-stage estimates



Note: Data from the Chicago Police Department. Each point represents an estimate of $(\hat{\beta}_1, \hat{\alpha}_1)$ from equation 9.

Table 1: Policing strategies used by surveyed departments

Strategy	Fraction using
Community policing	93.7%
Problem-oriented policing	88.9%
Hot spot policing	79.9%
Targeting specific problem addresses	91.5%
Targeting known offenders	79.3%

Source: Future Trends in Policing. 2014. Police Executive Research Forum. https://www.policeforum.org/assets/docs/Free_Online_Documents/Leadership/futuretrendsinpolicing2014.pdf

Table 2: RIFLE test results: 50 largest cities, 2010–2019

Outcome	p-value
Crime	
Homicides	0.17
Violent index crimes	0.02
Enforcement activities	
All arrests	0.74
Narcotics arrests	0.76
Enforcement harms	
Civilian-police killings	0.05

Note: Data from UCR and Fatal Encounters. Departments serving the 50 largest jurisdictions based on median population in 2010-2019. Details of the RIFLE analysis are described in section 2.1.

Table 3: RIFLE test results: Chicago police district commanders

Outcome	p-value
Crime	
Shootings	0.04
Violent felonies	0.51
Enforcement activities	
All arrests	0.34
Misdemeanor arrests	0.42
Narcotics arrests	0.50
Gun arrests	0.08
Street stops	0.01
Traffic stops	0.01
Enforcement harms	
Use of force incidents	0.07

Note: Data from the Chicago Police Department. Details of the RIFLE analysis are described in section 2.1.

Table 4: Estimated effects of the SDSCs: 7th and 11th districts, short-run

District	Outcome	Rate per 100K (post-treatment)		Difference		p-value		q-value
		Actual	Counterfactual	Units	%	w/o Resampling	w/ Resampling	
		(1)	(2)	(3)	(4)	(5)	(6)	(7)
7	Shooting Victims	13.1	34.6	-13.1	-62.1	0.095	0.098	0.109
11	Shooting Victims	19.4	43.4	-17.3	-55.3	0.083	0.085	0.109
7	Violent Felonies	206.8	220.4	-8.3	-6.2	0.667	0.673	0.811
11	Violent Felonies	214.8	196.6	13.2	9.3	0.200	0.224	0.596
7	Part 1 Crimes	482.6	600.6	-71.9	-19.7	0.095	0.023	0.160
11	Part 1 Crimes	525.3	552.5	-19.6	-4.9	0.429	0.334	0.716
7	All Crimes	1523.2	1557.5	-20.9	-2.2	0.750	0.848	0.811
11	All Crimes	1914.1	1718.6	141.1	11.4	0.167	0.221	0.596
7	Gun Arrests	32.8	31.0	1.1	5.9	0.952	0.855	0.854
11	Gun Arrests	30.5	20.9	6.9	45.9	0.167	0.269	0.395
7	Warrant Arrests	114.9	92.2	13.8	24.6	0.167	0.118	0.309
11	Warrant Arrests	135.8	99.2	26.4	36.9	0.200	0.080	0.249
7	Drug Arrests	46.0	36.0	6.1	27.8	0.619	0.477	0.589
11	Drug Arrests	206.5	191.4	10.9	7.9	0.667	0.650	0.705
7	Misdemeanor Arrests	288.9	312.7	-14.5	-7.6	0.333	0.233	0.395
11	Misdemeanor Arrests	350.7	279.4	51.4	25.5	0.048	0.005	0.031
7	All Arrests	584.3	605.5	-12.9	-3.5	0.667	0.599	0.705
11	All Arrests	827.5	656.6	123.3	26.0	0.200	0.005	0.031
7	Traffic Stops	3730.8	3145.2	356.8	18.6	0.417	0.367	0.473
11	Traffic Stops	2431.1	1806.3	450.8	34.6	0.143	0.074	0.249
7	Uses Of Force	37.8	29.9	4.8	26.2	0.381	0.376	0.602
11	Uses Of Force	27.7	36.2	-6.1	-23.3	0.500	0.374	0.602

Note: Data from the Chicago Police Department. Table reports estimated effects of the SDSCs in the 7th and 11th districts on crime and police activity outcomes in February 2017. Sharpened two-stage q-values controlling the FDR calculated using the procedure described in Benjamini et al. (2006) and implemented by Anderson (2008). For details, see section 4.2.

Table 5: Estimated effects of the SDSCs: Tier 1

	% Difference				
Outcome	1 Month	2 Months	3 Months	6 Months	11 Months
Shooting Victims	-31.9	-24.8	-20.7	-2.9	1.4
	[0.011]	[0.011]	[0.028]	[0.704]	[0.829]
	{0.012}	{0.012}	{0.030}	{1.000}	{1.000}
Violent Felonies	-9.6	-10.9	-7.9	-2.4	-7.6
	[0.019]	[0.026]	[0.091]	[0.507]	[0.042]
	{0.060}	{0.026}	{0.032}	{1.000}	{0.146}
Part 1 Crimes	-6.0	-7.2	-6.9	-1.8	-2.3
	[0.427]	[0.037]	[0.006]	[0.430]	[0.382]
	{0.167}	{0.026}	{0.009}	{1.000}	{0.342}
All Crimes	-5.3	-7.9	-5.9	-2.3	-3.3
	[0.088]	[0.003]	[0.006]	[0.530]	[0.248]
	{0.097}	{0.009}	{0.009}	{1.000}	{0.330}
Gun Arrests	15.6	9.7	16.9	18.2	18.5
	[0.239]	[0.328]	[0.046]	[0.168]	[0.001]
	{0.379}	{1.000}	{0.266}	{0.390}	{0.004}
Warrant Arrests	14.2	3.6	-0.9	5.3	9.6
	[0.275]	[0.743]	[0.908]	[0.332]	[0.090]
	{0.379}	{1.000}	{1.000}	{0.507}	{0.137}
Drug Arrests	4.1	13.8	17.3	71.7	58.0
	[0.808]	[0.199]	[0.070]	[0.001]	[0.001]
	{0.701}	{1.000}	{0.266}	{0.008}	{0.004}
Misdemeanor Arrests	8.7	-0.9	2.2	-0.7	4.0
	[0.037]	[0.805]	[0.439]	[0.920]	[0.704]
	{0.127}	{1.000}	{0.813}	{0.706}	{0.410}
All Arrests	11.4	5.7	4.7	8.3	9.4
	[0.031]	[0.402]	[0.448]	[0.345]	[0.194]
	{0.127}	{1.000}	{0.813}	{0.507}	{0.220}
Traffic Stops	4.3	-5.2	2.5	32.7	30.4
	[0.680]	[0.608]	[0.742]	[0.140]	[0.251]
	{0.701}	{1.000}	{1.000}	{0.390}	{0.220}
Uses Of Force	7.3	-6.3	6.1	5.9	6.2
	[0.709]	[0.772]	[0.635]	[0.481]	[0.330]
	{1.000}	{1.000}	{1.000}	{0.929}	{0.494}

Note: Data from the Chicago Police Department. Table reports estimated effects of the SDSCs in the Tier 1 districts as a whole, over different time intervals starting from February 2017. p-values calculated using the resampling method described in section 4.1 reported in square brackets. Sharpened two-stage q-values controlling the FDR calculated using the procedure described in Benjamini et al. (2006) and implemented by Anderson (2008) reported in curly brackets. Families of outcome variable used to calculate q-values are defined within each time interval. For details, see section 4.3.

Table 6: Estimated effects of HunchLab on officer time in a box, and of officer time on ShotSpotter alerts, by district

	Officer minutes			ShotSpotter alerts		
	Control mean	Effect of	First-stage	Control mean		Effect of minute
District	(analysis sample)	showing box	F-statistic	(full sample)	(analysis sample)	of officer time
	(1)	(2)	(3)	(4)	(5)	(6)
Pooled	10.05	0.2004**	6.16	0.0081	0.0130	-0.0029
		(0.0808)				(0.0044)
						[-0.0249,0.0191]
6	9.20	0.2675	1.03	0.0007	0.0012	0.0002
		(0.2633)				(0.0023)
7	11.48	0.4252**	6.51	0.0144	0.0183	-0.0093
		(0.1667)				(0.0061)
						[-0.0384,0.0199]
9	8.00	0.2866**	4.47	0.0043	0.0085	-0.0048
		(0.1355)				(0.0054)
						[-0.0548,0.0453]
10	8.93	0.1473	0.70	0.0009	0.0021	-0.0007
		(0.1761)				(0.0059)
11	10.28	0.1327	0.61	0.0212	0.0302	0.0079
		(0.1697)				(0.0291)
15	14.50	-0.2591	0.53	0.0129	0.0156	-0.0119
		(0.3574)				(0.0198)

Note: Data from the Chicago Police Department. First and second stage estimates from equation 9. Heteroskedasticity-robust standard errors clustered at the box level reported in parentheses. tF-adjusted 95% confidence intervals calculated using the method described in Lee et al. (2021) reported in brackets where first-stage F-statistic exceeds 4. See section 5 for details. * $p < 0.1$ ** $p < 0.05$ *** $p < 0.01$.