How Many Complaints Against Police Officers Can Be Abated by Incapacitating A Few "Bad Apples?"*

Aaron Chalfin[1] and Jacob Kaplan[2]

[1]University of Pennsylvania
[2]Princeton University

Abstract

Research Summary: The notion that the unjustified use of force by police officers is concentrated amongst a few "bad apples" is a popular descriptor which has gained traction in scholarly research and achieved considerable influence among policymakers. But is removing the bad apples likely to have an appreciable effect on police misconduct? Leveraging a simple policy simulation and data from the Chicago Police Department, we estimate that removing the top 10 percent of officers identified based on *ex ante* risk and replacing them with officers drawn from the middle of the risk distribution would have led to only a 4-6 percent reduction in use of force incidents in Chicago over a ten-year period.

Policy Implications: Our analysis suggests that surgically removing predictably problematic police officers is unlikely to have a large impact on citizen complaints. By assembling some of the first empirical evidence on the likely magnitude of incapacitation effects, we provide critical support for the idea that early warning systems must be designed, above all, to deter problematic behavior and promote accountability.

Keywords: police use of force, early warning systems, complaints against police

*For helpful comments, we thank Bocar Ba, Richard Berk, Maria Cuellar, Zubin Jelveh, Greg Ridgeway, Sarah Tahamont, and Peanut. Correspondence: Aaron Chalfin, Department of Criminology, 558 McNeil Building, 3718 Locust Walk, University of Pennsylvania, Philadelphia, PA 19104. E-Mail: achalfin@sas.upenn.edu.

1 Introduction

The idea that a small number of "bad apples" are responsible for an outsize share of complaints against police officers has gained considerable traction over the course of the last four decades both in the scholarly literature (Berkow, 1996; Alpert and MacDonald, 2001; Walker et al., 2001; Rozema and Schanzenbach, 2019; Goncalves and Mello, 2020) and in government reports (Christopher, 1991; Mollen, 1994) and popular media accounts (Arthur, 2018; Invisible Institute, 2018; Wu, 2019; Ba and Rivera, 2020; Kelly and Nichols, 2020; MacDonald and Klick, 2020). Such a claim is inspired by numerous anecdotal descriptions of "rogue cops" (Greek, 2007; Sherman, 2020), by research which documents the salience of social networks amongst police officers (Ouellet et al., 2019; Wood et al., 2019; Zhao and Papachristos, 2020) as well as by the Pareto principle (also sometimes called the "80/20 rule"), the empirical regularity that, in many areas of human inquiry, approximately 80 percent of the effects accrue from 20 percent of the causes.[1] With respect to policing, appeals to the Pareto principle were instrumental in spurring the creation of the first "early warning systems" to identify problematic police officers in the 1970s (Walker et al., 2000) and has informed numerous police reform initiatives in the intervening years (Alpert and Walker, 2000; Walker et al., 2001; Hughes and Andre, 2007). Given the increasing availability of "big data" to inform police practice (Ridgeway, 2018), many observers, including famously, *Campaign Zero*, have expressed optimism that large administrative datasets can potentially be put to use to identify problematic police officers and incapacitate them before they have the opportunity to do serious harm to members of the communities they serve (Sherman, 2018).

Empirically the claim that a small number of police officers account for an outsize share of serious misconduct rests on analyses of individually identified microdata on complaints against police officers. By

[1] The Pareto Principle derives from the observation of 19th century economist, Vilfredo Pareto, that, in many societies, 80 percent of the wealth is concentrated in the hands of 20 percent of the population (Pareto et al., 1971). In mathematical statistics, the Pareto principle has stimulated the study of "power law distributions" which characterize an astonishing array of natural phenomena in physics, biology, earth and planetary sciences as well as in the social and computational sciences (Newman, 2005).

collapsing the data at the officer level and sorting the officers in descending order with respect to the number of complaints they have generated, researchers can compute the share of complaints over a given time period that are accounted for by the top k percent of officers. Prior analyses from police departments across the United States suggest that a small share of officers account for a large share of complaints against police. Indeed a common estimate is that the top 2 percent of officers account for approximately 50 percent of known misconduct by police officers (Walker et al., 2001). As the other 98 percent of officers are responsible for the remaining 50 percent of misconduct, the implication is that the top 2 percent of officers are, incredibly, 49 times more likely to commit misconduct than other officers.[2] With respect to public policy, such an analysis suggests that if only the small number of "bad apples" can be identified and successfully intervened upon, law enforcement agencies could make substantial progress in reducing police misconduct without making any other institutional changes to policy or practice. Indeed, a relative risk ratio of 49 naively suggests that nearly 50 percent of use of force complaints could potentially be abated by replacing the top 2 percent of officers with officers drawn from the remainder of the distribution.[3]

Unfortunately, such an analysis suffers from three problems. First, the above computation assumes that we can predict bad acts among police officers with perfect foresight. Second, the analysis makes no provision for the replacement of "bad apples" with other police officers, who while less likely to commit misconduct, will nevertheless continue to generate complaints. Finally, the computation suffers from a simple but, to date, seldom identified problem which we refer to as "data density bias." Put simply, when the number of complaints is relatively small compared to the number of police officers, it will be true, by definition, that a small share of the officers will account for a large share of the known incidents. To see this, consider the simplified but nevertheless instructive case in which there are only 5 serious complaints

[2]To see this, consider a municipal law enforcement agencies which employs 1,000 officers and experiences 100 misconduct complaints. In this agency, the top 20 officers account for 50 complaints and the remaining 980 officers account for 50 complaints. The relative risk ratio is given by: $\frac{50/20}{50/980} = 49$.

[3]Suppose a department has N = 1,000 officers and m = 100 use of force complaints over some time period. The top 20 officers are known to account for 50 percent of the 100 complaints and the remaining 980 officers account for the other 50 percent of complaints. Replacing the top 20 officers with 20 officers whose risk is equal to the remainder of officers would lead to a reduction of 100 - 1000 $\times \frac{50}{980} = 51$ complaints.

filed in a city which employs 1,000 police officers. It is easy to see that even if all 5 complaints implicate a different officer, 100 percent of the complaints would be accounted for by just $\frac{5}{1{,}000} = 0.5$ percent of officers. While the headline — 0.5 percent of officers account for all of the serious complaints — sounds impressive, it is merely a statistical artifact that is intrinsic to the analysis of sparse data.[4] As we show in Section 2.1, sparse data are common in studies which use complaint data. Accordingly, calculations such as the one above have the potential to distort the policy conversation to a considerable degree.

How can we correct for data density bias? The solution lies in identifying a valid benchmark against which to compare a conventional assessment of the degree to which complaints are concentrated among police officers. In particular, we need to know what share of a law enforcement agency's use of force complaints would be accounted for by the top k percent of officers in a hypothetical world in which the use of force by police officers were completed unconcentrated.[5] Happily, such a counterfactual is easy to both identify and to compute. By randomizing complaints with replacement to police officers, we can generate a null distribution — the share of complaints that would be accounted for by the top k percent of officers in the complete absence of concentration. Referring to the example above, were we to randomize 5 complaints among 1,000 officers a large number of times, in nearly all iterations, the 5 complaints would be randomly assigned to different officers. Since 0.5 percent of officers will have accounted for 100 percent of the complaints in both the real data and the simulated data, we would conclude that there is, in fact, no concentration in the use of serious force by police officers. Accordingly, the Pareto principle would constructively fail to hold even though it would be supported by a naive analysis of the data. Put differently, some of the "bad apples" may have simply been "unlucky apples."[6]

[4] A similar argument has been made in the literature on crime concentration by Hipp and Kim (2017), Levin et al. (2017) and Chalfin et al. (2020) among others.

[5] We note that such a counterfactual is used to derive a null distribution of complaints and is not offered as a descriptor for how complaints are likely to be distributed in practice. Indeed, due to organizational priorities, officer self-sorting and peer networks (Ouellet et al., 2019; Wood et al., 2019; Zhao and Papachristos, 2020), there are strong reasons to believe that complaints will be concentrated — to some degree — among officers. The key question is how to accurately characterize the degree to which complaints are concentrated in the data.

[6] This conceptualization of risk is similar in spirit to an approach that is found Ridgeway and MacDonald (2009) who identify NYC police officers who are the most likely to engage in biased policing. In order to ensure that officers flagged by their statistical algorithm are, in fact, high-risk and not merely "unlucky," they motivate a statistical framework in which

To the extent that use of force is relatively unconcentrated, this narrows the scope for incapacitating problematic police officers to have a large effect on use of force complaints. However, a more salient policy question is how much force can be abated by incapacitating predictably problematic police officers, a question which hinges on the ability of analysts to make an *ex ante* prediction about which officers will receive future complaints. As we demonstrate in Section 2.3.1, even though there is reasonably strong persistence in use of force complaints throughout an officer's career, incapacitating the small number of officers who generate the greatest number of complaints early in their career is likely to lead to only a modest reduction in future use of force complaints. Drawing on a simple but realistic policy simulation, we estimate that replacing the 10 percent of officers who generated the largest number of use of force complaints early in their career with officers drawn from the middle of the distribution would have led to only a 4-6 percent reduction in use of force complaints against the Chicago Police Department over a ten-year period. The modesty of this impact is, in part, due to the difficulty of predicting future complaints and, in part, due to the extent to which data density bias has obfuscated the true degree of concentration in the use of force by police officers.

Our conclusion is that while incapacitating predictably problematic officers serves an important instrumental purpose, this practice is, in of itself, unlikely to lead to a large reduction in use of force complaints, absent appreciable deterrence or spillover effects or broader cultural change. As such, early warning systems should be designed to promote accountability among a broader set of officers, rather than to serve as a narrowly-tailored tool to surgically remove high-risk personnel. While the importance of accountability has long been a focal point in the scholarly literature on early warning systems (Alpert and Walker, 2000; Walker et al., 2001), references to the concentration of misconduct amongst a small number of "bad apples" are pervasive in popular media accounts and public commentary. By assembling some of the first empirical evidence on the likely magnitude of incapacitation effects, we provide critical support for the

the risk threshold is raised until false discovery rates are tolerably low.

idea that early warning systems must be designed, above all, to deter problematic behavior and promote accountability.

2 Empirical Example

2.1 Data and Methods

We explore the extent to which complaints are concentrated amongst police officers using individually identified microdata made public by the Chicago Police Department. These data come from the Invisible Institute's Citizens Police Data Project which is a collection of nearly 250,000 complaints against Chicago Police Department officers filed since 1988 (Ba and Rivera, 2020). The data which were collected via a lengthy and painstaking series of Freedom of Information Act requests have been used in recent research on police use of force and police-civilian interactions including that of Ba and Rivera (2019), Holz et al. (2019), Faber and Kalbfeld (2019), Rozema and Schanzenbach (2019), Ouellet et al. (2019), Wood et al. (2019), Ba et al. (2020), Rim et al. (2020), Rim et al. (2020), and Zhao and Papachristos (2020) among others.[7] This dataset is ideal for our purposes as it includes information on all complaints as well as a full roster of Chicago police officers including, critically, those who have never been named in a complaint. The data also includes information on each officer's assigned police command, his or her race and gender as well as tenure within the department.

In this research, our primary focus is on citizen complaints, which implicate one or more Chicago police officers. Naturally, not every complaint will be justified and, in practice, many complaints will fail to be sustained upon detailed review. Likewise, some errant behavior, in particular criminal acts, may be committed by officers while they are off-duty (Fyfe, 1980; Kane and White, 2009). While we acknowledge that complaints are an imperfect proxy for official misconduct, we believe a focus on complaints is both appropriate and useful for three reasons. First, research by Rozema and Schanzenbach (2019) has shown that

[7]The data were downloaded from their website https://beta.cpdp.co/

complaints are a surprisingly good predictor of high-impact events such as lawsuit payouts by municipal officials. While legal settlements are rare and therefore extraordinarily difficult to predict, complaints are more common and, as such, have greater predictive signal. Second, by focusing on all complaints instead of sustained complaints or use of force incidents recorded via police department record keeping, we use data that has not been filtered through the lens of what a law enforcement agency deems problematic and which therefore may better reflect community norms. Finally, by focusing on all complaints instead of sustained complaints, we generate a lower bound on the extent to which data density bias distorts the policy conversation. Since sustained complaints are a subset of all complaints, the degree of data density bias will be even greater in such an analysis. Given that citizen complaints are an incomplete measure of use of force by police officers, we also analyze internal use of force data captured in the Chicago Police Department's "tactical response reports" (TRR). These data also come from the Invisible Institute's Citizens Police Data Project. These reports are filled out by police officers who have used force in the course of their duties and accordingly the data provide an alternative measure of use of force events which include incidents that are not captured in complaint data.[8]

To explore the concentration of complaints, we focus on complaints made against Chicago police officers recorded by the police during the five-year period between September 17, 2012 and September 17, 2017.[9] For use of force incidents recorded by the "tactical response reports", we use the five-year window between April 12, 2011 and April 12, 2016 as this is the final date of available TRR data. For each complaint, we have information on the police officers involved in the complaint, the police district to which they are assigned, their race and gender as well as the nature of the complaint. We focus on the $11,283$ police

[8]Per departmental directives, a tactical response report is generally required under the following circumstances: 1) a subject is injured or alleges injury resulting from an officer's use of force, 2) force is used to subdue a suspect who physically attacks an officer or who threatens to use violence, 3) an officer employs force that is reasonably expected to cause pain or injury (even if it does not, in fact, result in injury or complaint of injury), 4) an officer uses a "less-lethal weapon" or 5) the officer uses deadly force including discharging a firearm, effecting a head blow, using a chokehold or using force that resulted in a hospital admission. A tactical response report is generally not required for the use of escort holds and pressure compliance techniques which do not result in an injury or allegation of injury or for the use of joint manipulation techniques, wristlocks, armbars or other firm grips utilized in conjunction with handcuffing and searching techniques which do not result in an injury or allegation of injury. For further details, see http://directives.chicagopolice.org/forms/CPD-11.377.pdf.

[9]September 17, 2017 is the last date where the data indicates that a police officer retired.

officers who were employed by the Chicago Police Department as of September 17, 2017 — the last date for which we have information on complaints — and use the last five years of available data so that we have up to five years of data for each officer.[10] Among these 11,283 officers, between September 17, 2012 and September 17, 2017, there were 17,247 complaints of which 2,885 involved the use of force. Thus, over a five-year period, there were approximately 1.5 total complaints and 0.2 use of force complaints per police officer. During the five-year period ending on April 12, 2016, there were 24,217 tactical response reports involving force filed by CPD officers, or approximately 2.1 per officer.

In a second analysis, we focus on the cohort of Chicago police officers hired between 2000 and 2007 and follow them prospectively over the following ten years.[11] For this analysis we limit police officers to those whose first reported unit is one of Chicago's police districts.[12] This excludes approximately seven percent of officers in this sample who either are in a non-district unit (e.g. a special unit such as the "special functions support unit") or whose data on their starting unit is not available. We rank officers with respect to the number of complaints they receive early in their careers and use this information in order to predict future complaint risk. This exercise forms the basis for our policy simulation, described in Section 2.3, in which we identify high-risk officers using data generated during their early career probationary period and simulate the replacement of these officers using a variety of different heuristic approaches that simulate how replacement might happen. We use this policy simulation to estimate the share of citizen complaints over a ten-year period that could have been abated solely by incapacitating the "bad apples", identified *ex ante*.

[10] We restrict the data to individuals employed at the rank of police officer, police officer-training officer, sergeant or police-officer-detective in order to focus on the subset of officers who routinely encounter citizens while on patrol.

[11] Since tactical response reports are not available prior to 2004, in these analysis we focus on the cohort of officers hired between 2004 and 2007 and follow them over a slightly shorter post-period.

[12] In cases where the first reported unit for the officer is the training unit, we use the second reported unit.

2.2 How Concentrated are Citizen Complaints?

In **Figure 1**, we explore the concentration of overall complaints against Chicago police officers (Panel A), use of force complaints specifically (Panel B) and tactical response reports generated by CPD officers (Panel C). In each graph, there are three lines. The solid red line plots the cumulative distribution of complaints — that is, the share of complaints that are accounted for by the top k percent of officers. The dashed gray line plots the null distribution, the cumulative distribution function which arises from randomly assigning complaints to police officers, with replacement. The solid black line is a 45 degree line which represents uniformity in concentration — that is, the condition in which top k percent of officers account for k percent of complaints for all values of k. Naturally, uniformity can only hold prior to the saturation point at which 100 percent of the complaints are accounted for. The figure can be used to make a number of useful comparisons which reveal the extent to which complaints are concentrated amongst police officers. Graphically, the degree of data density bias is greatest when the simulated density function under randomization lies closer to the empirical density function than to the 45 degree line. Indeed when the empirical and simulated density functions lie on top of one another, complaints are, in fact, unconcentrated.

Referring to Panel A, we see that the top 20 percent of officers, ranked according to the number of complaints they have generated, account for approximately 65 percent of the complaints and the top half of officers account for nearly all of the complaints. However, the slope of the curve is steepest at the top of the distribution. Here, we observe that the top 2 percent of police officers account for approximately 14.2 percent of total complaints against the department. Put differently, the top 2 percent of officers are $\frac{\frac{14.2}{2}}{\frac{(100-14.2)}{(100-2)}} = 8.1$ times more likely to generate complaints than the remaining 98 percent of officers. This naive computation suggests that complaints are concentrated to a large degree amongst a very small number of officers. However, this computation does not account for the obfuscating effect of data density bias. To see how important data density bias is empirically, we turn to our simulated data in which we

randomly assigned complaints to police officers with replacement. Referring to the dashed gray line, under random assignment, the top 2 percent of officers generate 6.9 percent of the complaints. This suggests that, even under random assignment, the top 2 percent of officers are $\frac{\frac{6.9}{2}}{\frac{(100-6.9)}{(100-2)}} = 3.6$ times more likely to generate complaints than the remaining 98 percent of officers. As such, in the real-world data, the top 2 percent of officers are $\frac{8.1}{3.6} = 2.3$ times more likely to generate complaints than in a condition in which there is no concentration in use of force by construction. While this computation suggests that complaints are, in fact, concentrated, the naive comparison overstates the degree of concentration by a factor of nearly 4.

Next, we turn to use of force complaints which account for 17 percent of all complaints against Chicago Police officers during our five-year study period. Given that these types of complaints are less common, the degree to which data density bias obfuscates comparisons will be greater. In Panel B, we see that the top 10 percent of officers, ranked according to the number of complaints they have generated, account for 70 percent of the complaints and the top 16 percent of officers account for all of the use of force complaints. In other words, 84 percent of officers generated no use of force complaints during the sample period. At the top of the distribution, the top 2 percent of officers account for 26.1 percent of the use of force complaints. In other words, these officers are $\frac{\frac{26.1}{2}}{\frac{(100-26.1)}{(100-2)}} = 17.3$ times more likely to generate complaints than the remaining 98 percent of officers. This comparison suggests an extraordinary degree of concentration and accordingly that the Chicago Police Department could appreciably reduce use of force complaints by removing a small number of "bad apples." However, the figure also shows a considerable degree of concentration even when complaints are randomized to officers. Indeed, even in the simulated data, the top 2 percent of officers account for 16.6 percent of use of force complaints. As such, even under randomization, these officers are $\frac{\frac{16.6}{2}}{\frac{(100-16.6)}{(100-2)}} = 9.8$ times as likely to generate force complaints than other officers. Thus, rather than use of force complaints being 17 times more common among the top 2 percent officers, they are, in fact, only $\frac{17.3}{9.8} = 1.8$ times as likely once data density bias is accounted for.

A similar story holds for tactical response reports for which the top 2 percent of officers account for

17.2 percent of the reports. Given that in the simulated data, the top 2 percent of officers account for 6 percent of reports, once data density bias is accounted for, the top 2 percent of officers are, in fact, only 3.3 times as likely to be involved in use of force reports than other officers. While this finding confirms that use of force is substantively concentrated and therefore that either individual characteristics, spatial variation in complaint risk or network effects are important drivers of use of force incidents, use of force is considerably less concentrated than the standard metrics imply.

2.3 Policy Simulation

2.3.1 Persistence in Complaints

Although complaints are not nearly as concentrated as they might have seemed at first blush, there still appear to be officers who are at an elevated risk to generate citizen complaints. We next assess the extent to which "bad apples" are *ex ante* predictable. We begin by considering the extent to which there is persistence in the generation of complaints and the application of force against suspects. Specifically, we assess the degree to which the officers who are most likely to generate complaints early in their careers are also the most likely to generate complaints later in their careers. Next, we motivate a simple but informative policy simulation in which we estimate the share of use of force complaints which could be abated by removing a small number of officers who generate the greatest number of complaints early in their careers and replacing them with officers identified using a variety of different heuristics.

In settings in which there are rich cross-sectional data — for example, detailed demographic data or pre-employment information — predictions about police officer risk are typically made using sophisticated machine learning-based algorithms (Ridgeway and MacDonald, 2009; Carton et al., 2016; Chalfin et al., 2016; Helsby et al., 2018) or, at a minimum, logistic regression (Leinfelt, 2005; White, 2008).[13] The

[13]For example, leveraging a wealth of data on pre-employment characteristics, Chalfin et al. (2016) use stochastic gradient boosting to predict police misconduct among a sample of police officers in Philadelphia. Machine learning-based algorithms are indeed used in a variety of settings in the criminal justice system including to inform decisions about sentencing (Berk and Hyatt, 2015), parole (Berk, 2017) and arraignment (Berk et al., 2016; Kleinberg et al., 2018). For an excellent reference on the development of machine learning algorithms in the U.S. criminal justice system, we refer readers to Berk (2019).

advantage of machine learning methods in such a context is that the approach allows researchers to automate the detection of signal in the data, a task which is complicated considerably when the number of predictors is large and the relationships between variables are non-linear and conditional (Hastie et al., 2009).[14]

Given the longitudinal nature of complaint data, we focus instead on a simpler but, we argue, especially policy-relevant prediction exercise which captures the extent to which there is persistence in complaints among officers. We focus on the eight cohorts of Chicago police officers hired between 2000 and 2007 and who remain employed by the Chicago police department in 2017.[15] For each officer, we retain 11.5 years of data and divide the 11.5-year sample period into an 18-month pre-period and a ten year post-period.[16] We choose a pre-period of 18 months as this is the standard probationary period for new police officers hired in Chicago — as a robustness check, we repeat this analysis using a five-year probationary period. The purpose of this exercise is to identify the police officers who generate the largest number of complaints during their probationary period and to see how many of them continue to generate an outsize number of complaints throughout the prime of their careers. While there are a number of reasons to expect that complaints will be serially correlated, persistence in complaints might also vary over the life course, thus making ex ante prediction more challenging. Reasons for this include changing life circumstances (Linn, 2009) as well as the phenomenon of "late career misconduct" (Harris, 2011, 2016; Stinson, 2020) among others. Overall, the ability of an early warning system to incapacitate a meaningful share of misconduct will depend on the degree to which there is temporal persistence in the generation of complaints, given the many external factors are at play.

A first-order issue in carrying out such an analysis concerns whether or not to condition on the police

[14] See Beutler et al. (1985), Leinfelt (2005) and Ridgeway (2020) for assessments of the predictors of police use of force and Lum (2016) and Ridgeway (2016) for assessments with respect to race.

[15] Naturally over a ten-year period, some officers will be terminated as a result of a use of force incident. However, termination is exceedingly rare — less than 0.2 percent of officers are terminated annually (Ba and Rivera, 2020).

[16] Given the smaller time period available in the tactical response reports data, we use an eight-year post-period for this data.

command to which an officer is assigned. Given that crime, organizational culture, officer characteristics and peer networks vary between commands, it stands to reason that any analysis that does not account for an officer's command may lead to misleading downstream estimates. In order to account for the fixed properties of commands and all of the attendant ways in which these shape officer behavior, in our main analyses, we condition on an officer's assigned command. In particular, within each Chicago district, among the officers who are in the top k percent, ranked according to the number of complaints accrued during their probationary period, we identify the share who are in the top j percent of officers, ranked according to use of force complaints in the ten-year post-period.

The estimand we report corresponds with the "positive predictive value" in the prediction literature.[17]

The results of this exercise are presented in **Table 1** which reports positive predictive values for k, j = 2, 5, 10 and 20 percent.

Among the officers in the top 2 percent of complaints in the probationary period, 1.96 percent are also in the top 2 percent in the post-period and nearly one quarter are in the top 10 percent in the post-period. Among the officers in the top 5 percent of complaints in the probationary period, 22 percent are in the top 10 percent in the post-period. With respect to use of force complaints, 3.92 percent of officers who are in the top 2 percent of the pre-period distribution are also in the top 2 percent in the post-period distribution, 17.7 percent are in the top 10 percent in the post-period distribution and 37.3 percent are in the top 20 percent in the post-period distribution. With respect to tactical response reports for which there are denser data, 16.7 percent of officers who are in the top 2 percent of the pre-period distribution are also in the top 2 percent in the post-period distribution, 44.4 percent are in the top 10 percent in the post-period distribution and over 60 percent are in the top 20 percent in the post-period distribution.[18]

[17] Formally, the positive predictive value is computed as the number of true positives (here, those who are in the top j percent of the use of force distribution in the post-probationary period) divided by the sum of the number of true positives and false positives.

[18] When we use city-wide data to identify "bad apples" and do not condition on an officer's assigned district, the estimates are substantively similar, albeit slightly larger, reflecting the fact that prediction is yet more difficult among a group of officers who are likely to be sorted on individual characteristics and who are likely to face a more similar crime environment. These estimates are presented in **Appendix Table 1**.

2.3.2 Simulating the Replacement of "Bad Apples"

Overall, Table 1 suggests that, even within districts, there is considerable persistence in complaints over an officer's career thus generating optimism that complaints are a predictable phenomenon. However, the critical policy question is how many complaints could be abated by terminating high-risk officers identified on the basis of their early career complaint activity and replacing them with less risky officers. In order to answer this question, we construct a simple but informative policy simulation in which "bad apples" identified ex ante are replaced with officers who have a lower likelihood of generating complaints. Recognizing that there is considerable uncertainty in how "bad apples" might be replaced and by whom, we employ a wide range of heuristics each of which is informed by a large and rapidly proliferating literature on police organization (Faber and Kalbfeld, 2019; Ouellet et al., 2019; Wood et al., 2019; Zhao and Papachristos, 2020).

We begin with a baseline assumption that "bad apples" are replaced by the median officer which we conceptualize by replacing removed officers with officers who are randomly drawn from the middle 20 percent of the distribution of officers when ranked according to complaints in the 1.5 year probationary period. Given recent research which suggests that there is a strong degree of homophily within a police department with respect to characteristics such as race, gender, tenure and district assignment (Wood et al., 2019), the proposition that a high-risk officer will be replaced with a median officer is a strong assumption. Accordingly, we relax this assumption by allowing high-risk officers to be replaced using a variety of heuristics, each of which is intended to capture officer sorting in a different way. First, instead of replacing "bad apples" with the median officer, we instead perform replacement using the top 10-30 percent (the 70th to 90th percentile) of officers ranked according to prior complaints. This analysis accounts for the possibility that officers might, in fact, be replaced by officers with similar underlying propensities to generate complaints. Second, recognizing that officers may be non-randomly sorted into districts either due to self-selection or organizational practices which route officers with certain characteristics to certain districts, we instead replace the removed officers with officers drawn at random from

the middle 20 percent of the distribution *among officers who work in the same district.* Third, we re-run our analysis using the alternative assumption that "bad apples" are replaced with the top 10-30 percent (the 70th to 90th percentile) of officers ranked according to prior complaints and who work in the same district.

Fourth, recognizing the importance of sorting with respect to race, gender and tenure as well as network links along these characteristics (Wood et al., 2019), we re-run the analysis assuming that "bad apples" are replaced with an officer randomly drawn from the 40th-60th percentile of same district-race bin, noting that by focusing on cohorts of officers hired during the same year, we are already capturing the effects of sorting by officer tenure.[19] Fifth, we re-run the analysis assuming that "bad apples" are replaced with those drawn from the middle 20 percent of the distribution of officers in the same district-gender bin as the removed officer. Finally, we replace the "bad apples" with officers drawn from the same district-race-gender bin, removing the requirement that the officer is also in the middle 20 percent of the district-race or district-gender distribution.

The results of this analysis are presented in **Table 2** which reports the share of post-period complaints that would have been abated if the top k percent of officers identified at the end of the probationary period were replaced with an equivalent number of officers drawn using one of the heuristics described above. For instance, for $k = 2$ percent, we remove the 61 police officers who generated the greatest number of complaints during the 18-month probationary period and replace those officers with 61 officers drawn using one of the above heuristics. In the table, Panel A presents results for the standard 18-month probationary period; in Panel B we consider an alternative probation period of 5 years.

Using our baseline heuristic, we estimate that removing the top 2 percent of officers — identified ex ante — from circulation would abate just 0.89 percent of total complaints, 1.19 percent of use of force complaints and 1.74 percent of use of force incidents as proxied by tactical response reports. Even the

[19]We assign officers to three race groups: Black, white and other. In the event that no such officer is available, we assign a replacement on the basis of race and district alone. This is rarely the case though.

replacement of the top 10 percent of the workforce — an enormously difficult task given the current rate of termination of approximately 0.2 percent per year (Ba and Rivera, 2020) — with officers drawn from the middle of the distribution is estimated to reduce total complaints and use of force complaints by just 4.6 and 6.1 percent, respectively.[20]

Importantly, while there is uncertainty about precisely how "bad apples" would be replaced in a real-world setting, the table shows that the estimated incapacitation effects of an early warning system are not very sensitive to the heuristic employed. Naturally, to the extent that high-risk officers are replaced with others who are similar with respect to an underlying propensity to generate complaints, the magnitude of the incapacitation effects will be smaller. Across each of the seven heuristics employed, we estimate that removing the top 2 percent of officers would have incapacitated between 0.8 and 1.2 percent of use of force complaints. Similarly, we project that removing the top 10 percent of officers would have incapacitated between 4.9 and 6.3 percent of use of force complaints. While our estimates imply that incapacitating the "bad apples" is unlikely to lead to a large-scale reduction in complaints, we pause here to point out that in a large city like Chicago, even a small proportional reduction, can mean hundreds fewer complaints annually. As research by Wood et al. (2020) suggests that a 6 percent decline in complaints leads to the abatement of approximately $20 million in lawsuit settlements, even small declines can have important fiscal implications as well as implications for police-community relations which often hinge on a small number of events that are especially salient or well-publicized.

Perhaps an 18-month probationary period is insufficient to be able to identify high-risk officers. In order to assess the sensitivity of our policy simulation to this parameter of the analysis, we repeat this exercise focusing on a longer probationary period. Estimates using a five-year probationary are presented in Panel B of Table 2. While using a longer probationary period improves the quality of our predictions,

[20] In **Appendix Table 2**, we use tactical response reports, which are denser, to predict use of force complaints. Identifying "bad apples" ex-ante using the tactical response reports, we project that replacing the top 2 percent of "bad apples" would have incapacitated between 1.8 and 2.3 percent of use of force complaints and that replacing the top 10 percent of "bad apples" would have incapacitated between 7.3 and 9.3 percent of use of force complaints. These estimates are slightly larger than those which use prior use of force complaints suggesting that the denser data are helpful, albeit modestly.

there are two key drawbacks of such an approach. First, it is preferable to identify high-risk officers early in their careers before they have the opportunity to generate complaints. Second, it is generally more difficult to terminate or reassign officers after their official probationary period has ended. When we use a five-year probationary period, we estimate that terminating the top 2 percent of officers, ranked according to use of force complaints in the five-year pre-period, would have abated just 3.6-4.6 percent of use of force complaints thus suggesting that even a much longer probationary period does not solve the problem. However, using a five-year probationary period, the benefits to terminating the top 10 percent of officers are considerably greater. While terminating such a large number of officers who have already served five years on the force would likely be politically challenging, this analysis highlights the importance making predictions using denser data.[21]

Given that an accounting of the raw data suggest that the top 2 percent of officers are 17 times more likely to generate use of force complaints than other officers, these estimates — which are fairly modest — may appear surprising. However, the estimates are, in fact, sensible given that future complaints cannot be predicted with perfect foresight and that complaints are not especially concentrated among a small number of officers. For example, officers in the 90th percentile of the probationary period distribution of use of force are not very different in the number of complaints generated (2.76 per officer during the ten-year post-period) than the median officer (1.67 per officer during the ten-year post-period). In the next section, we discuss the implication of these findings for public policy.

3 Policy Implications

In thinking about complaints, there are four possible ways of describing the ease with which problematic police officers can be identified and incapacitated. The most helpful scenario is when complaints are both highly concentrated and highly predictable. When this is the case, policymakers will find it easy to identify

[21] In **Appendix Table 3** we replicate Table 2 ignoring district assignment information. Estimated incapacitation effects, as expected, are slightly larger.

the "bad apples" and likewise achieve large reductions in complaints by incapacitating those officers. A second possibility is that complaints are predictable but not particularly concentrated. In this scenario, it is possible to identify which officers will commit bad acts but, given that there are likely to be feasibility constraints with respect to the number of officers who can be incapacitated, the number of complaints that can be abated will be tempered by the lack of appreciable concentration in the data. A third possibility is that complaints are concentrated but not very predictable. Such a scenario might come to pass if, in a given year, a large share of the complaints accrues to a small share of officers but, in each year, the problematic officers are different. Such a scenario is unwelcome in the sense that a shifting policy environment makes it difficult to successfully intervene prior to the accrual of harms. A fourth possibility is that complaints are neither particularly predictable nor particularly concentrated. This possibility leaves little room for optimism that complaints can be meaningfully reduced solely through prediction and incapacitation.

The data suggest that the use of force by police officers is concentrated amongst a small number of problematic officers to a degree, albeit far less concentrated than naive calculations would suggest. This finding, in turn, suggests that while the scope for incapacitating problematic police officers to have an appreciable effect on misconduct is narrower than the standard calculations imply. Such a claim is further underscored by the difficulty of predicting who the most problematic police officers will be at the time they are hired or early in an officer's career (Cuttler and Muchinsky, 2006; Fyfe and Kane, 2006; White, 2008; Chalfin et al., 2016). Consistent with the prior literature, when we use an officer's early career accumulation of complaints to predict an his or her subsequent career performance, the accumulation of complaints is somewhat predictable. However, the positive predictive value is modest — between 2 and 39 percent depending on the threshold used. Accordingly, the number of false positives remains high, complicating the extent to which such a process could be used to make personnel decisions such as terminating or even simply reassigning police officers (Goldman and Puro, 2001; White, 2008; Dharmapala et al., 2019). While the number of false positives can be reduced by raising the risk threshold used to flag problematic officers

(Ridgeway and MacDonald, 2009), the cost of doing so is inevitably fewer abated complaints — a tradeoff that is informed by the impossibility of simultaneously minimizing both Type I and Type II errors.

With respect to public policy, the most direct implication of this analysis is that, absent appreciable deterrence effects or broader cultural change, early warning systems that are designed to identify problematic police officers and incapacitate them — either through termination or re-assignment — are unlikely to lead to large reductions in the use of force. Likewise, a surgical focus on "bad apples" may be less effective than broad-based measures to improve managerial practices and increase accountability (Sherman, 1978; Skolnick and Fyfe, 1993; Ivkovic, 2009; Mummolo, 2018). While pitting these two policy solutions against each other, in principle, presents a false choice, in practice, constraints on political capital may require policymakers to invest in a limited set of actions. With respect to the efficacy of broad-based police reform efforts, while there continues to be a dearth of high-quality evidence in this domain (Sherman, 2018; Engel et al., 2020), there is, at least, some evidence to support the efficacy of de-escalation training (Engel et al., 2020) and procedural justice training (Owens et al., 2018; Nagin and Telep, 2020; Wood et al., 2020), federal oversight of police agencies (Powell et al., 2017; Goh, 2020) as well as the use of and training in non-lethal weapons (MacDonald et al., 2009; Sousa et al., 2010). There is likewise support for the idea that reforms involving police unions may be effective (Dharmapala et al., 2019) especially if unions can be incentivized to "self-regulate" which might potentially be encouraged by transferring the burden of liability insurance from municipalities to unions (Ramirez et al., 2018). Finally, as noted by Mummolo (2018), police officers tend to be highly responsive to managerial directives, leaving room for optimism that procedural reforms can dramatically alter officer behavior.

A second implication of this analysis is that it is critical for policymakers to incentivize better reporting and discovery of police misconduct (Long et al., 2013; Knox et al., 2019). Incomplete reporting of misconduct by citizens inevitably leads to noisy data which, in turn, leads to poor predictability (Ivkovic, 2009) and a diminishing of the ability of data-driven early warning systems to have maximum impact. Happily,

there is evidence that more complete reporting can be achieved through actions that are available to many municipal policymakers. For example, Ba (2018) finds that when police departments make it more difficult for citizens to report complaints, the number of complaints decreases. As such, there is evidence that citizens may be responsive to lowering the cost of reporting officer misconduct. Likewise, as suggested by Rozema and Schanzenbach (2019), requiring civilians making allegations to swear out an affidavit before an investigation may proceed may have a chilling effect on reporting misconduct. Similarly, to the extent that there are other markers of complaints such as internal investigations, ad hoc performance assessments or pre-employment information, this information will be critical to deploy in order to further enhance the predictability of bad acts. One especially promising idea is the collection of customer service data arising from police-citizen encounters (Burn, 2010). While there are challenges to collecting data from individuals who are the recipients of police service, the richness of such data might well be a goldmine for prediction, especially over a short time window.

Finally, while our policy simulation suggests that identifying and surgically incapacitating the "bad apples" is unlikely to have a large and direct impact on use of force, early warning systems, coupled with rigorous oversight and genuine accountability have the potential to have a far larger effect by generating deterrence or spillover effects, or by holistically changing departmental culture. While we are unable to comment directly on the magnitude of deterrence effects, recent research by Quispe-Torreblanca and Stewart (2019) sheds light on the potential size of network spillovers. Using data from the London Metropolitan police force and variation induced by the movement of officers across the organization, the authors find that a 10 percent increase in misconduct committed by an officer's peers, increases an officer's future misconduct by 8 percent. What do these results — which suggest that the magnitude of spillovers is proportional to the incapacitation of peer misconduct — mean for our analysis? The implication is that while removing a large number of officers could generate substantively important network spillovers, removing a small number of bad apples is unlikely to generate appreciable spillover effects. To see this, assume that consistent with

Quispe-Torreblanca and Stewart (2019), for every 1 percent of misconduct that is incapacitated through the removal of "bad apples," we can expect an additional 0.8 percent of misconduct to be removed via network spillovers thus increasing the effect size by 80 percent. While this large multiplier is consistent with meaningful and behaviorally-important peer network effects, since the incapacitation effects are so small in the first place, even when a large spillover effect is assumed, the share of complaints that we project would be abated by incapacitating the top 2 percent and 10 percent of officers, respectively, would be just 1.4 percent and 10 percent, respectively.

To the extent that an early warning system serves as a deterrent to officers on the margin, these efforts may well be capable of producing marked changes in use of force well in excess of the estimates we report in this research. We therefore emphasize that our policy simulation, by design, does not identify the promise of early warning systems more generally or what the effects might be at scale (Sampson et al., 2013).Indeed the net impact of data-driven efforts to identify "bad apples" will depend critically on the extent to which these efforts are coupled with initiatives that change behavior among police officers who are unlikely to be flagged as being high-risk.

References

Alpert, G. P. and J. M. MacDonald (2001). Police use of force: An analysis of organizational characteristics. *Justice Quarterly 18*(2), 393–409.

Alpert, G. P. and S. Walker (2000). Police accountability and early warning systems: Developing policies and programs. *Justice Research and Policy 2*(2), 59–72.

Arthur, R. (2018). 130 chicago officers account for 29 percent of police shootings. *The Intercept*.

Ba, B. (2018). Going the extra mile: The cost of complaint filing, accountability, and law enforcement outcomes in chicago. Technical report, Working paper.

Ba, B., D. Knox, J. Mummolo, and R. Rivera (2020). Diversity in policing: The role of officer race and gender in police-civilian interactions in chicago. Technical report, Working Paper. https://scholar. princeton. edu/sites/default/files/jmummolo ....

Ba, B. and R. Rivera (2020). Police think they can get away with anything. that's because they usually do.

Ba, B. A. and R. Rivera (2019). The effect of police oversight on crime and allegations of misconduct: Evidence from chicago. *Faculty Scholarship at Penn Law*. (19-42).

Berk, R. (2017). An impact assessment of machine learning risk forecasts on parole board decisions and recidivism. *Journal of Experimental Criminology 13*(2), 193–216.

Berk, R. (2019). *Machine learning risk assessments in criminal justice settings*. Springer.

Berk, R. and J. Hyatt (2015). Machine learning forecasts of risk to inform sentencing decisions. *Federal Sentencing Reporter 27*(4), 222–228.

Berk, R. A., S. B. Sorenson, and G. Barnes (2016). Forecasting domestic violence: A machine learning approach to help inform arraignment decisions. *Journal of Empirical Legal Studies 13*(1), 94–115.

Berkow, M. (1996). Weeding out problem officers. *Police Chief 63*, 21–29.

Beutler, L. E., A. Storm, P. Kirkish, F. Scogin, and J. A. Gaines (1985). Parameters in the prediction of police officer performance. *Professional Psychology: Research and Practice 16*(2), 324.

Burn, C. (2010). The new south wales police force customer service programme. *Policing: A Journal of Policy and Practice 4*(3), 249–257.

Carton, S., J. Helsby, K. Joseph, A. Mahmud, Y. Park, J. Walsh, C. Cody, C. E. Patterson, L. Haynes, and R. Ghani (2016). Identifying police officers at risk of adverse events. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 67–76.

Chalfin, A., O. Danieli, A. Hillis, Z. Jelveh, M. Luca, J. Ludwig, and S. Mullainathan (2016). Productivity and selection of human capital with machine learning. *The American Economic Review 106*(5), 124–27.

Chalfin, A., J. Kaplan, and M. Cuellar (2020). Measuring marginal crime concentration: A new solution to an old problem.

Christopher, W. (1991). Independent commission on the los angeles police department.(1991) report of the independent commission on the los angeles police department. *Los Angeles, CA: The Commission*.

Cuttler, M. J. and P. M. Muchinsky (2006). Prediction of law enforcement training performance and dysfunctional job performance with general mental ability, personality, and life history variables. *Criminal Justice and Behavior 33*(1), 3–25.

Dharmapala, D., R. H. McAdams, and J. Rappaport (2019). Collective bargaining and police misconduct: Evidence from florida.

Engel, R. S., H. D. McManus, and T. D. Herold (2020). Does de-escalation training work? a systematic review and call for evidence in police use-of-force reform. *Criminology & Public Policy*.

Faber, J. W. and J. R. Kalbfeld (2019). Complaining while black: Racial disparities in the adjudication of complaints against the police. *City & Community 18*(3), 1028–1067.

Fyfe, J. J. (1980). Always prepared: Police off-duty guns. *The Annals of the American Academy of Political and Social Science 452*(1), 72–81.

Fyfe, J. J. and R. Kane (2006). *Bad cops: A study of career-ending misconduct among New York City police officers*. John Jay College of Criminal Justice.

Goh, L. S. (2020). Going local: Do consent decrees and other forms of federal intervention in municipal police departments reduce police killings? *Justice Quarterly*, 1–30.

Goldman, R. L. and S. Puro (2001). Revocation of police officer certification: A viable remedy for police misconduct. *Saint Louis University Law Journal 45*, 541.

Goncalves, F. and S. Mello (2020). A few bad apples?: Racial bias in policing. *The American Economic Review*.

Greek, C. (2007). The big city rogue cop as monster: Images of nypd and lapd. *Monsters in and among us: Toward a Gothic criminology*, 164–198.

Harris, C. (2011). Problem behaviors in later portions of officers' careers. *Policing: Nn International Journal of Police Strategies & Management*.

Harris, C. J. (2016). Towards a career view of police misconduct. *Aggression and Violent Behavior 31*, 219–228.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Helsby, J., S. Carton, K. Joseph, A. Mahmud, Y. Park, A. Navarrete, K. Ackermann, J. Walsh, L. Haynes, C. Cody, et al. (2018). Early intervention systems: Predicting adverse interactions between police and the public. *Criminal Justice Policy Review 29*(2), 190–209.

Hipp, J. R. and Y.-A. Kim (2017). Measuring crime concentration across cities of varying sizes: Complications based on the spatial and temporal scale employed. *Journal of Quantitative Criminology 33*(3), 595–632.

Holz, J., R. Rivera, and B. A. Ba (2019). Spillover effects in police use of force. *U of Penn, Inst for Law & Econ Research Paper* (20-03).

Hughes, F. and L. Andre (2007). Problem officer variables and early-warning systems. *Police Chief 74*(10), 164.

Invisible Institute, T. (2018). The citizens police data project.

Ivkovic, S. K. (2009). Rotten apples, rotten branches, and rotten orchards: A cautionary tale of police misconduct. *Criminology & Public Policy 8*, 777.

Kane, R. J. and M. D. White (2009). Bad cops: A study of career-ending misconduct among new york city police officers. *Criminology & Public Policy 8*(4), 737–769.

Kelly, J. and M. Nichols (2020). We found 85,000 cops who've been investigated for misconduct. now you can read their records. *USA Today*.

Kleinberg, J., H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan (2018). Human decisions and machine predictions. *The Quarterly Journal of Economics 133*(1), 237–293.

Knox, D., W. Lowe, and J. Mummolo (2019). Administrative records mask racially biased policing. *The American Political Science Review*, 1–19.

Leinfelt, F. H. (2005). Predicting use of non-lethal force in a mid-size police department: A longitudinal analysis of the influence of subject and situational variables. *The Police Journal 78*(4), 285–300.

Levin, A., R. Rosenfeld, and M. Deckard (2017). The law of crime concentration: An application and recommendations for future research. *Journal of Quantitative Criminology 33*(3), 635–647.

Linn, E. (2009). *Arrest Decisions: What Works for the Officer?* Number 5. Peter Lang.

Long, M. A., J. E. Cross, T. O. Shelley, and S. Kutnjak Ivković (2013). The normative order of reporting police misconduct: Examining the roles of offense seriousness, legitimacy, and fairness. *Social Psychology Quarterly 76*(3), 242–267.

Lum, C. (2016). Murky research waters: The influence of race and ethnicity on police use of force. *Criminology & Public Policy 15*, 453.

MacDonald, J. and J. Klick (2020). Hire more cops to reduce crime. *City Journal*.

MacDonald, J. M., R. J. Kaminski, and M. R. Smith (2009). The effect of less-lethal weapons on injuries in police use-of-force events. *American Journal of Public Health 99*(12), 2268–2274.

Mollen, M. (1994). *Commission report*. The Commission.

Mummolo, J. (2018). Modern police tactics, police-citizen interactions, and the prospects for reform. *The Journal of Politics 80*(1), 1–15.

Nagin, D. S. and C. W. Telep (2020). Procedural justice and legal compliance: A revisionist perspective. *Criminology & Public Policy*.

Newman, M. E. (2005). Power laws, pareto distributions and zipf's law. *Contemporary Physics 46*(5), 323–351.

Ouellet, M., S. Hashimi, J. Gravel, and A. V. Papachristos (2019). Network exposure and excessive use of force: Investigating the social transmission of police misconduct. *Criminology & Public Policy 18*(3), 675–704.

Owens, E., D. Weisburd, K. L. Amendola, and G. 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.

Pareto, V. et al. (1971). Manual of political economy.

Powell, Z. A., M. B. Meitl, and J. L. Worrall (2017). Police consent decrees and section 1983 civil rights litigation. *Criminology & Public Policy 16*(2), 575–605.

Quispe-Torreblanca, E. G. and N. Stewart (2019). Causal peer effects in police misconduct. *Nature Human Behaviour 3*(8), 797–807.

Ramirez, D., M. Wraight, L. Kilmister, and C. Perkins (2018). Policing the police: Could mandatory professional liability insurance for officers provide a new accountability model. *American Journal of Criminal Law 45*, 407.

Ridgeway, G. (2016). Officer risk factors associated with police shootings: a matched case–control study. *Statistics and Public Policy 3*(1), 1–6.

Ridgeway, G. (2018). Policing in the era of big data.

Ridgeway, G. (2020). The role of individual officer characteristics in police shootings. *The ANNALS of the American Academy of Political and Social Science 687*(1), 58–66.

Ridgeway, G. and J. M. MacDonald (2009). Doubly robust internal benchmarking and false discovery rates for detecting racial bias in police stops. *Journal of the American Statistical Association 104*(486), 661–668.

Rim, N., B. Ba, and R. Rivera (2020). Disparities in police award nominations: Evidence from chicago. In *AEA Papers and Proceedings*, Volume 110, pp. 447–51.

Rim, N., R. Rivera, A. Kiss, B. Ba, et al. (2020). The black-white recognition gap in award nominations. Technical report.

Rozema, K. and M. Schanzenbach (2019). Good cop, bad cop: Using civilian allegations to predict police misconduct. *American Economic Journal: Economic Policy 11*(2), 225–68.

Sampson, R. J., C. Winship, and C. Knight (2013). Translating causal claims: Principles and strategies for policy-relevant criminology. *Criminology & Public Policy 12*, 587.

Sherman, L. W. (1978). *Scandal and reform: Controlling police corruption.* University of California Press.

Sherman, L. W. (2018). Reducing fatal police shootings as system crashes: Research, theory, and practice.

Sherman, L. W. (2020). Targeting american policing: Rogue cops or rogue cultures?

Skolnick, J. H. and J. J. Fyfe (1993). *Above the law: Police and the excessive use of force.* Free Press New York.

Sousa, W., J. Ready, and M. Ault (2010). The impact of tasers on police use-of-force decisions: Findings from a randomized field-training experiment. *Journal of Experimental Criminology 6*(1), 35–55.

Stinson, P. M. (2020). *Criminology explains police violence*, Volume 1. University of California Press.

Walker, S., G. P. Alpert, and D. J. Kenney (2000). Early warning systems for police: Concept, history, and issues. *Police Quarterly 3*(2), 132–152.

Walker, S., G. P. Alpert, and D. J. Kenney (2001). *Early warning systems: Responding to the problem police officer.* US Department of Justice, Office of Justice Programs, National Institute of Justice.

White, M. D. (2008). Identifying good cops early: Predicting recruit performance in the academy. *Police Quarterly 11*(1), 27–49.

Wood, G., D. Roithmayr, and A. V. Papachristos (2019). The network structure of police misconduct. *Socius 5*, 2378023119879798.

Wood, G., T. R. Tyler, and A. V. Papachristos (2020). Procedural justice training reduces police use of force and complaints against officers. *Proceedings of the National Academy of Sciences 117*(18), 9815–9821.

Wu, K. J. (2019). Study finds misconduct spreads among police officers like contagion. *PBS*.

Zhao, L. and A. V. Papachristos (2020). Network position and police who shoot. *The ANNALS of the American Academy of Political and Social Science 687*(1), 89–112.

Figure 1: Actual Versus Simulated Concentration of Complaints Against Chicago Police Officers

A: All Complaints

B: Use of Force Complaints

C: Tactical Response Reports

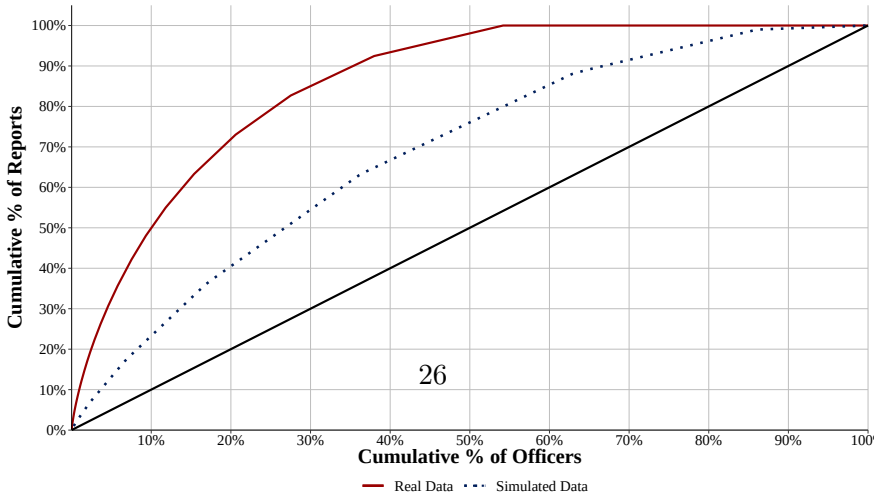

Table 1: Persistence in Complaints or Tactical Response Reports For Top Officers Within Each Police District

	Top 2%	Top 5%	Top 10%	Top 20%
Top 2%	1.96	13.73	23.53	39.22
Top 5%	4.23	12.68	21.83	36.62
Top 10%	3.06	8.50	17.01	30.27
Top 20%	2.66	6.99	14.81	26.29

(a) **Panel A: All Complaints**

	Top 2%	Top 5%	Top 10%	Top 20%
Top 2%	3.92	7.84	17.65	37.25
Top 5%	3.52	11.97	22.54	35.21
Top 10%	3.40	8.5	17.01	28.23
Top 20%	3.00	8.32	14.81	25.62

(b) **Panel B: Use of Force Complaints**

	Top 2%	Top 5%	Top 10%	Top 20%
Top 2%	16.67	33.33	44.44	61.11
Top 5%	8.20	26.23	37.70	60.66
Top 10%	4.38	15.33	25.55	49.64
Top 20%	3.55	10.64	19.50	36.52

(c) **Panel C: Tactical Response Report (TRR)**

Note: For both total complaints (Panel A) and use of force complaints (Panel B), we estimate the percentage of officers who are in the top k percent of complaints in their probationary period (the first 18 months after they are hired) that are also in the top k percent of complaints in the 10 year follow-up period. For tactical response reports (Panel C), we estimate the percentage of officers who are in the top k percent of complaints in their probationary period (the first 18 months after they are hired) that are also in the top k percent of complaints in the 8 year follow-up period. The rows denote the probationary period share while the columns denote the share in the follow-up period. "Bad apples" are identified within an officer's assigned district.

Table 2: Policy Simulation: Estimated Percent Change in Complaints When Replacing the Top k Percent of Officers Within Each Police District With Other Officers

	Top 2%			**Top 10%**		
Replacement Group	All Complaints	Use of Force Complaints	Tactical Response Report (TRR)	All Complaints	Use of Force Complaints	Tactical Response Report (TRR)
Citywide 40-60th	-0.89	-1.19	-1.74	-4.56	-6.14	-9.95
Citywide 70-90th	-0.82	-1.00	-1.58	-4.61	-5.68	-9.43
Within-District 40-60th	-0.88	-1.20	-1.61	-4.48	-6.28	-9.24
Within-District 70-90th	-0.73	-0.94	-1.43	-4.01	-5.15	-8.43
Within-District 40-60th, Same Race	-0.79	-1.11	-1.63	-4.72	-6.36	-9.36
Within-District 40-60th, Same Gender	-0.80	-0.99	-1.46	-4.23	-5.91	-8.82
Within-District Same Race and Gender	-0.70	-0.79	-1.34	-3.70	-4.99	-8.88

(a) **Top k Percent of Officers for 18-Month Probationary Period**

Replacement Group	All Complaints	Use of Force Complaints	Tactical Response Report (TRR)	All Complaints	Use of Force Complaints	Tactical Response Report (TRR)
Citywide 40-60th	-3.07	-4.56	-3.00	-13.81	-16.80	-17.24
Citywide 70-90th	-2.28	-3.65	-2.20	-11.26	-13.30	-13.51
Within-District 40-60th	-2.97	-4.51	-2.99	-13.46	-16.67	-17.15
Within-District 70-90th	-2.20	-3.59	-2.14	-9.93	-12.33	-12.18
Within-District 40-60th, Same Race	-3.06	-4.51	-2.88	-13.65	-17.33	-17.28
Within-District 40-60th, Same Gender	-2.92	-4.38	-2.87	-13.39	-16.56	-16.91
Within-District Same Race and Gender	-2.67	-3.97	-2.55	-12.35	-15.65	-15.97

(b) **Top k Percent of Officers for 5-Year Probationary Period**

Note: For total complaints, use of force complaints and tactical response reports, we estimate the reduction in the number of complaints that would have accrued if a given share, k, of Chicago police officers, ranked according to the number of complaints they accrued among officers in their district during their probationary period had been terminated at the end of their probationary period and replaced with other officers using a variety of replacement approaches. Panel A considers an 18-month probationary period and a 10-year post-period (8-year post-period for tactical response reports); Panel B considers a five-year probationary period and a 6-year post-period (4.5 year post-period for tactical response reports).

ONLINE APPENDIX

Appendix Table 1: Persistence in Complaints or Tactical Response Reports For Top Officers Citywide

	Top 2%	Top 5%	Top 10%	Top 20%
Top 2%	6.56	11.48	32.79	50.82
Top 5%	5.26	9.87	25.66	42.11
Top 10%	3.93	10.16	21.64	39.34
Top 20%	3.28	8.20	16.07	30.33

(a) **Panel A: All Complaints**

	Top 2%	Top 5%	Top 10%	Top 20%
Top 2%	4.92	18.03	27.87	49.18
Top 5%	5.26	15.79	25.66	42.11
Top 10%	4.59	11.15	19.02	32.46
Top 20%	3.61	9.02	17.70	30.49

(b) **Panel B: Use of Force Complaints**

	Top 2%	Top 5%	Top 10%	Top 20%
Top 2%	25.00	39.29	60.71	71.43
Top 5%	15.28	26.39	43.06	61.11
Top 10%	9.72	16.67	29.86	52.08
Top 20%	6.23	11.42	20.42	37.37

(c) **Panel C: Tactical Response Report (TRR)**

Note: For both total complaints (Panel A) and use of force complaints (Panel B), we estimate the percentage of officers who are in the top k percent of complaints in their probationary period (the first 18 months after they are hired) that are also in the top k percent of complaints in the 10 year follow-up period. For tactical response reports (Panel C), we estimate the percentage of officers who are in the top k percent of complaints in their probationary period (the first 18 months after they are hired) that are also in the top k percent of complaints in the 8 year follow-up period. The rows denote the probationary period share while the columns denote the share in the follow-up period. "Bad apples" are identified using citywide data, without conditioning on an officer's assigned district.

Appendix Table 2: Policy Simulation: Estimated Percent Change in Complaints When Replacing the Top k Percent of Officers Within Each Police District With Other Officers

	Top 2%	**Top 10%**
Replacement Group	Use of Force Complaints	Use of Force Complaints
Citywide 40-60th	-2.23	-9.25
Citywide 70-90th	-1.99	-8.13
Within-District 40-60th	-2.02	-8.23
Within-District 70-90th	-1.80	-7.30
Within-District 40-60th, Same Race	-1.90	-8.25
Within-District 40-60th, Same Gender	-1.84	-7.56
Within-District Same Race and Gender	-1.95	-8.36

(a) **Top k Percent of Officers for 18-Month Probationary Period**

Replacement Group	Use of Force Complaints	Use of Force Complaints
Citywide 40-60th	-2.51	-19.61
Citywide 70-90th	-1.83	-16.54
Within-District 40-60th	-2.43	-19.23
Within-District 70-90th	-1.63	-14.90
Within-District 40-60th, Same Race	-2.14	-19.07
Within-District 40-60th, Same Gender	-2.24	-19.06
Within-District Same Race and Gender	-1.89	-18.05

(b) **Top k Percent of Officers for 5-Year Probationary Period**

Note: We estimate the reduction in the number of use of force complaints that would have accrued if a given share, k, of Chicago police officers, ranked according to the number of tactical response reports (TRR) they accrued among officers in their district during their probationary period, had been terminated at the end of their probationary period and replaced with other officers using a variety of replacement approaches. Panel A considers an 18-month probationary period and a 8-year post-period; Panel B considers a five-year probationary period and a 4.5-year post-period.

Appendix Table 3: Policy Simulation: Estimated Percent Change in Complaints When Replacing the Top k Percent of Officers Citywide With Other Officers

	Top 2%			**Top 10%**		
Replacement Group	All Complaints	Use of Force Complaints	Tactical Response Report (TRR)	All Complaints	Use of Force Complaints	Tactical Response Report (TRR)
Citywide 40-60th	-1.41	-1.97	-3.58	-6.20	-7.18	-11.29
Citywide 70-90th	-1.31	-1.73	-3.30	-5.83	-6.28	-10.17
Within-District 40-60th	-1.33	-2.00	-3.39	-6.24	-7.35	-10.7
Within-District 70-90th	-1.05	-1.48	-3.09	-7.24	-7.23	-11.43
Within-District 40-60th, Same Race	-1.26	-1.88	-3.25	-6.45	-7.39	-10.64
Within-District 40-60th, Same Gender	-1.23	-1.74	-3.27	-6.08	-6.97	-10.37
Within-District Same Race and Gender	-1.03	-1.43	-3.10	-5.46	-5.97	-10.37

(a) **Top k Percent of Officers for 18-Month Probationary Period**

Replacement Group	All Complaints	Use of Force Complaints	Tactical Response Report (TRR)	All Complaints	Use of Force Complaints	Tactical Response Report (TRR)
Citywide 40-60th	-4.00	-6.07	-5.16	-15.07	-18.48	-18.51
Citywide 70-90th	-3.06	-5.03	-3.93	-11.21	-14.03	-13.29
Within-District 40-60th	-3.83	-5.93	-5.09	-14.80	-18.45	-18.46
Within-District 70-90th	-2.90	-4.82	-3.86	-13.88	-16.04	-15.87
Within-District 40-60th, Same Race	-3.87	-6.07	-5.09	-15.04	-19.08	-18.78
Within-District 40-60th, Same Gender	-3.77	-5.79	-5.04	-14.76	-18.36	-18.37
Within-District Same Race and Gender	-3.44	-5.40	-4.78	-13.96	-17.36	-17.50

(b) **Top k Percent of Officers for 5-Year Probationary Period**

Note: For total complaints, use of force complaints and tactical response reports, we estimate the reduction in the number of complaints that would have accrued if a given share, k, of Chicago police officers, ranked according to the number of complaints they accrued citywide during their probationary period had been terminated at the end of their probationary period and replaced with other officers using a variety of replacement approaches. Panel A considers an 18-month probationary period and a 10-year post-period (6-year post-period for tactical response reports); Panel B considers a five-year probationary period and a 8-year post-period (4.5 year post-period for tactical response reports).