State Policies and Police Personnel Decisions*

Aaron Chalfin
University of Pennsylvania

Dylan Fitzpatrick
University of Chicago

Jens Ludwig
University of Chicago and NBER

December 14, 2020

Abstract

Reducing police misconduct is, in large part, about finding predictable *ex ante* risk and reducing it, through a combination of human resource policies and human capital development, as well as having well-functioning accountability and disciplinary systems. This paper discusses ways in which state policymakers can help encourage and support local police departments in these efforts. We begin by discussing what is known about the ability of new "big data" methods to capture predictable risk. We then discuss various ways in which state-level policies can be helpful in improving existing systems, and conclude with thoughts about the highest-priority topics where future research is most desperately needed.

*Our thanks to Jennifer Doleac and members of the Arnold Ventures policing roundtable for helpful discussion and suggestions. Please address correspondence to Jens Ludwig: jludwig@uchicago.edu.

1 Introduction

In the summer of 2020, millions of people marched all over the country demanding changes to policing, cheered on by a majority of their fellow Americans. The result has been an active debate about what the future of policing and public safety should look like in the United States. Survey data suggest the public is divided on questions like whether the police should be defunded or abolished. But surveys suggest there is much more consensus that America should strive to reduce the harms that arise from law enforcement, in whatever form law enforcement might take in the future. Of particular concern to the public is reducing the harms from police use of force against civilians, which is both far more common in the United States than in other developed nations and also disproportionately involves people of color. With public energy mobilized to address this issue, the next challenge is to identify constructive steps forward.

Reducing police misconduct is, in large part, about finding predictable *ex ante* risk and reducing it, through a combination of human resource policies and human capital development, as well as having well-functioning accountability and disciplinary systems. Police departments have a great deal of flexibility in setting the rules for how they select from their initial job applicant pools, and with respect to the decision to hire someone at the end of their probationary period. Risk of future misconduct, if it can be predicted, could be used to help inform those decisions. Once officers are officially hired, termination becomes a matter of a disciplinary system that is a great deal more proscribed and is inevitably *ex post*, making judgments after the fact about whether some police action is grounds for firing. But *ex ante* prediction of risk can still be useful for purposes of targeting supports, such as additional training or mental health treatment, as with the type of 'early warning system' that the US Department of Justice's Civil Rights division typically recommends as part of consent decrees with departments. So while some of the most important HR decisions come relatively early in the career cycle, our statistical analysis suggests that the ability to predict risk improves substantially over the course of an officer's career as more data on each officer is accumulated (at least given the type of data police departments currently collect).

In this paper we discuss what role states might play in such efforts. One of the challenges to large-scale changes in policing in the US is that law enforcement is, to a large extent, a local affair. Of the approximately 750,000 law enforcement officers employed in the U.S., more than 60% are employed by municipal agencies and another 25% are employed by county sheriff's departments.[1] While states have broad authority to shape local policing through their legislative, executive and judicial powers, the prospect of state involvement raises questions about which decisions should be made by which level of government. However there do appear to be some principles that can help guide decisions about what situations might benefit from involvement by the higher level of government (like states), such as:

[1] See: https://www.bjs.gov/content/pub/pdf/nsleed.pdf

- States can help local police departments overcome *fixed costs* or *economies of scale* with certain police reforms, such as the construction of predictive risk models and delivery or evaluation of officer supports and training
- States can help in dealing with *externalities* and *information failures* at the local level, as for example when fired police officers are hired at another department because of restrictions on the sharing of past employment records
- States can *redistribute resources*, analogous to the education domain where states provide an increasingly large share of K-12 spending to help overcome differences across localities in property tax bases

In the next section, we discuss what is currently known about how well we can predict an officer's future performance on the job at each stage of the HR pipeline (hiring, at the end of the probationary period, and once on the job). This requires, for starters, some way to define what we even mean by misconduct and performance on the job. The third section of the paper discusses different ways in which state government specifically might encourage or support local changes at each step of the HR pipeline.

2 Identifying High-Risk Officers

The idea that a small share of officers are responsible for an outsize share of complaints and misconduct has gained some traction over the course of the last four decades both in the scholarly literature (Alpert and MacDonald, 2001; Walker et al., 2001; Rozema and Schanzenbach, 2019) as well as in government reports (Christopher, 1991; Mollen, 1994) and popular media accounts (Arthur, 2018; Invisible Institute, 2018; Ba and Rivera, 2020; Kelly and Nichols, 2020; MacDonald and Klick, 2020). This led cities to pioneer the creation of data-driven "early warning systems" to identify high-risk officers, an innovation that departments began experimenting with more than 40 years ago (Walker et al., 2000). In this section, we discuss what is known about the ability to predict different outcomes at different stages of the officer's career.

There is a separate and equally important decision that we do not address in detail here, which is not just what outcomes society might wish to reduce, but which aspects of officer performance society wishes to actively encourage and select for in the HR pipeline. While previous social science research has relied on researcher judgment about what proxy measures to use to capture the idea of 'police productivity'[2], the definition of what society

[2]While early research tended to use arrests as a proxy for police productivity (Zhao et al., 2003), more recent research has focused on other potentially more salient dimensions of policing including engagement in activities such as business checks, community meetings, or investigative follow-ups with citizens (Wu and Lum, 2017), the rate at which arrests are validated by convictions (Ater et al., 2014; Chalfin and Goncalves, 2020; Weisburst, 2020), racial unbiasedness (Persico and Todd,

means by a 'good officer' is ultimately a normative one that needs to be defined by society itself through debate and the political process.

2.1 Predicting risk of what?

Police officers are trained to respond to suspects using force that is objectively reasonable and proportional to threats posed to officers and others. While the ultimate objective is to minimize unnecessary force, typically the outcome of focal concern is an illegal or (as departments call it) 'out of policy' use of force. One practical challenge in trying to predict this outcome is that, in most jurisdictions, the investigative process that determines whether a given incident in which an officer uses force is illegal or out of policy is slow. If the average investigation length is K, then a prediction model built at time T would only be able to draw on data from completed use-of-force investigations that run through time period T-K. Another challenge from a statistical perspective is that typically only a modest share of police uses of force are found to be illegal or out of policy. This means that even in large police departments, while there may be far too many such cases from a public policy perspective, from a narrow statistical perspective there may be too few such cases to enable the construction of an accurate predictive model.

In practice, this typically leads to the use of alternative proxy measures as the outcomes to be predicted. Figure 1 shows correlations between different officer-level event categories, using data from one urban police department in the United States. The figure includes events that vary in severity, ranging from serious adverse events (e.g., complaints of excessive force or complaints resulting in a suspension) to less severe outcomes which occur with much greater frequency (e.g., any citizen complaint). Also included are indicators of off-duty misconduct, which are relevant to departments under the hypothesis that officers who exhibit problematic behavior off the job, such as domestic violence or substance abuse, could, in principle, be at elevated risk for misconduct on the job.

One striking feature of the data shown in Figure 1 is how modest the correlations tend to be between many candidate measures of police misconduct. So policymakers must make some difficult decisions about which outcome to predict with statistical models, or whether to make the outcome some aggregate of multiple measures (and, if so, which ones).

Another policy question highlighted by Figure 1 is how to think about candidate measures of misconduct or risk of misconduct that are correlated with standard measures of 'output' that historically have been commonly used by police departments. For example the figure shows that outcomes like any complaint (whether sustained or not) or any use of force (whether found illegal or out of policy, or not) have among the highest correlations with the number of arrests an officer makes. In principle there could be correlations with other candidate measures of officer 'output' or 'productivity' as well; we focus narrowly on arrests

2006; Goncalves and Mello, 2020) and internal proxies of success using HR data (Sanders, 2008).

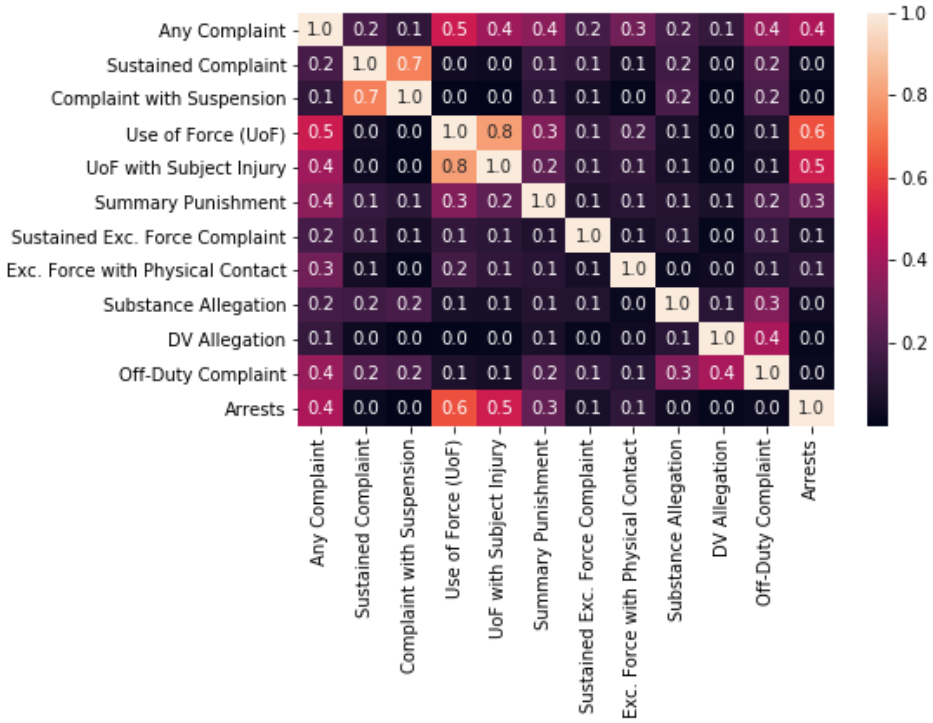

Figure 1: Correlations between officer-level adverse event categories and arrests.

here just for illustrative purposes, because that is the measure we happen to have access to. The interpretation of these correlations and their implications for police department practice will depend on (among other things) one's views of the public safety value of any particular 'output' measure, and hence the degree to which one is worried about dis-incentivizing the candidate measure of police output.

The above discussion suggests that social science and statistics have a contribution to make here, by highlighting the prevalence and inter-correlation of different candidate outcome measures, as well as the degree to which different outcomes can be predicted at different career stages. But ultimately, some judgment by policymakers themselves will be required to resolve ambiguities and uncertainties that the data cannot, and to take normative positions about tradeoffs that must be made in light of these uncertainties.

2.2 Assessing Risk During the Hiring Stage

Evaluating early-career hires presents a challenge in most occupations, and identifying good candidates can be particularly difficult for police departments due to the range of skills that officers must apply on a daily basis while on the job. Police are expected to be physically fit, to demonstrate emotional intelligence when handling people in crisis or taking reports from witnesses, and to maintain acute situational awareness during conflict. Many of these tasks are quite different from what people have done in their civilian lives.

While hiring requirements vary across departments, the process of applying to be a police officer typically involves a series of interviews and physical tests that are intended to evaluate the range of skills required of good police and to assess risk of behavioral issues. Applicants must first pass a written exam in order to proceed with the hiring process that assesses applicants' situational judgment, deductive and inductive reasoning skills, and basic reading and writing ability. Applicants also undergo a background check and a psychological evaluation, which usually involves taking a written personality test[3] and an interview with a trained psychologist intended to assess suitability for a career in law enforcement.

Research is surprisingly sparse on the question of how well current HR systems use information collected at the hiring stage to select officers at low risk for misconduct. Chalfin et al. (2016) explore this question, considering how the existing HR system's rank-ordering of applicants at one large urban police department in the US relates to risk of future officer misconduct, where the proxy measure for that outcome is likelihood that the officer would generate a physical abuse complaint. Replacing those at the bottom of the HR system's list (that is, bottom-decile applicants according to the existing HR system) with a better-ranked officer (from the middle segment of the HR system's rank-ordered list) would actually *increase* complaints by 0.6%. Put differently, the existing HR system seems to select officers at relatively *higher*, rather than *lower*, risk of generating a physical abuse complaint.

This tells us that police HR systems (or at least this department's HR system at the time these data were collected) are either selecting officers based on some other outcome besides risk of physical abuse, or else are not utilizing the information relevant to predicting physical abuse. For example, despite the breadth of information gathered on potential officers by assessment instruments during the application process, hiring decisions often rely heavily on applicants meeting minimum performance thresholds on the written and physical fitness exams. The physical exam presents a significant barrier to successful hire, even though mid-career officers list physical agility and fitness last among rankings of competencies important to policing (Bissett et al., 2012).[4] This has some analogues to research in education suggesting that existing public school systems do not put much weight on those few factors at the hiring stage that are, in fact, predictive of a teacher's effectiveness on the job.

A careful look at the same data from the application process as described above reveals that there is indeed some under-utilized 'signal' there about risk of future misconduct that could be utilized at the hiring stage. Chalfin et al. (2016) builds a data-driven statistical model to predict officer risk. One empirical challenge is that outcomes may be a function of not just person-level factors but situational factors as well; that is, officer assignments may affect the likelihood of the outcome and we do not wish to confuse the influence of person-

[3]See, for example: https://psychcentral.com/lib/minnesota-multiphasic-personality-inventory-mmpi/

[4]Physical exams may also enforce disparities in applicant success rates across certain demographic groups such as race and gender (Lonsway, 2003).

level with situational job-assignment factors, or what we call *task confounding*. To solve this problem Chalfin et al. (2016) focus in some of their analyses on only new cohorts of officers, who are often automatically assigned to higher-crime areas during their probationary-officer periods (so that all young officers are experiencing the same situation).

When Chalfin et al. (2016) rank-ordered newly-hired officers by their predicted *ex ante* risk derived from a statistical (machine learning) model, 22% of the highest-risk decile were later found to generated a physical abuse complaint. As a way to think about the implications for predictability, approximately 17% of all newly-hired officers go on to generate a complaint for physical abuse. So picking newly-hired officers at random would leave us with a subgroup that commits misconduct at this base rate. Equivalently, if the authors had built a statistical model that had no predictive power at all, the highest-risk officers according to that model would also commit misconduct at a rate of 17%. So the ratio of the rate at which predicted-high-risk officers have some outcome divided by the outcome's overall base rate is one measure of the model's predictive accuracy, known as *lift*. The lift in this case is equal to $\frac{22.4}{17} = 1.3$.

These estimates suggest that it would be possible to replace high-risk officers with those drawn from the middle of the predicted distribution and reduce physical abuse complaints by approximately 3%. This is obviously not a panacea by itself. Whether it is meaningfully large will depend in part on what achieving that 3% decline would cost, for example what would be required to expand the applicant pool by enough to not hire the highest-risk decile but still meet the department's overall hiring targets, and what the impact on use of force would be from other candidate interventions that could have been implemented with those resources instead.[5]

It may be possible to better predict officer risk at the hiring stage if police departments expanded the type of information that was collected as part of the hiring process. Recently developed assessment tools for hiring and recruiting could provide additional data points that could be tailored for police hiring to be predictive of later success on the job.[6] We return to this point below.

2.3 Assessing Risk At the End of the Probationary Period

Probationary review provides police departments with an opportunity to reassess an early career hire's suitability for the job based on performance data gathered over twelve to eighteen months on the job. For example, disciplinary actions such as suspensions or summary

[5]As another example, there is some evidence to suggest that scores on written psychological tests are correlated with indicators of job performance and risk of misconduct in police officers. For example, Roberts et al. (2019) find that officers with elevated scores on an aggressiveness scale are 3.6-3.9 times more likely to receive a poor job performance rating with respect to citizen relations, and officers with high scores on a cynicism scale are 4.0-5.2 times more likely to receive a poor rating of problem-solving and decision-making ability by a supervisor.

[6]See: https://www.pymetrics.ai/science

punishments, frequent incidents involving a subject injured by use of force, and citizen complaints all provide signals that may indicate an officer is facing difficulties in adjusting to the new job. The initial period of employment presents a chance to observe people carrying out the relevant set of tasks in the same or similar settings that they will be expected to work in for the remainder of their career.

The initial period of employment presents a chance to observe people carrying out the relevant set of tasks in the same or similar settings that they will be expected to work in for the remainder of their career. For example, Staiger and Rockoff (2010) find that releasing the bottom 10% of teachers based on performance in the first year on the job would result in an increase of roughly 0.01 standard deviations in student performance, corresponding to approximately 2-3% of the average improvement from one year of school (Aaronson et al., 2007).

A similar phenomenon seems to exist in policing. Chalfin and Kaplan (2020) use early career data collected during an officer's probationary period to simulate replacing officers with highest predicted risk with those drawn from the middle of the distribution of officers. The authors find that replacing the top 2% of officers in the predicted risk distribution with those from the middle of the risk distribution within their own district results in a 1-2% reduction in use of force complaints. The total number of use-of-force complaints averted will of course depend not just on their predictability, but on what share of retention decisions are altered. Replacing the top 10% of officers in the predicted risk distribution with officers drawn from the middle of the risk distribution is estimated to yield a 6-9% reduction in use of force complaints. Here again, such policies are far from a panacea for reducing adverse outcomes; whether the size of these gains should be considered meaningfully large or not depends on the costs of the policies that would be required to achieve them and what other policies might be foregone in order to enact them.

For the present paper we have also carried out some original data analysis using data from an urban police department in the U.S. and find that *aggregating outcomes* may improve predictability at this stage of the officer's career, and hence make risk mitigation policies more effective and more cost-effective. Predictive factors for this analysis included citizen complaints, use-of-force records, and summary punishment records indicating relatively minor disciplinary actions imposed by the department, defined as counts based on the specific type of event and how recently the event occurred. We employed gradient-boosted decision trees to predict whether a adverse event occurs in the two years following the point of prediction, for officers with eighteen months of experience or less as a sworn officer.

Prediction performance on several adverse event categories using data up to the end of the probationary period are reported in Table 1. When restricted to probationary period data, we see that prediction models do only slightly better than picking at random at identifying officers who will receive citizen complaints in the future. Officers in the top 1% of predicted risk for any citizen complaint went on to receive additional complaints at a rate of 75 per 100

Table 1: Performance of prediction models on predicting adverse events for early-career officers with 18 months or less on the job.

	Top 1%			Top 10%		
	Precision	Lift	Recall	Precision	Lift	Recall
Any citizen complaint	0.75	1.64	0.02	0.54	1.18	0.12
EF complaint with phys. contact	0.00	0.00	0.00	0.00	0.00	0.00
Complaint with suspension	0.00	0.00	0.00	0.00	0.00	0.00
Domestic violence complaint	0.00	0.00	0.00	0.00	0.00	0.00
Aggregate of adverse events	0.75	4.44	0.05	0.26	1.52	0.15

officers, and while this rate of predictive success may appear high, the relevant base rate of complaints against officers in this jurisdiction is 46 per 100. So if we take the share of flagged officers who experience the outcome (known as *precision*) and then divide that by the base rate, a ratio that tells us something about model predictive accuracy, this measure of so-called *lift* is 1.6. Put differently, our model is 1.6 times as good as flagging officers at random. The table also shows that rare outcomes pose particular challenges. Our models failed to predict rarer outcomes, such as complaints of excessive force with unnecessary physical contact, complaints resulting in suspension, and complaints of domestic violence. Of the highest 1% and 10% predicted-risk officers from these models, zero actually experienced the outcomes in the evaluation period.

Prediction models performed best on an aggregate outcome measure of adverse events, which combined off-duty complaints, suspensions, and sustained complaints. Officers at the top 1% of the predicted risk distribution again went on to experience an adverse event at a rate of 75 per 100 officers. This is the same rate at which the top 1% of predicted risk for any citizen complaint go on to experience complaints, but the base rate is much lower for this aggregate measure of potentially more serious adverse outcomes (17 per 100) compared to the base rate for all complaints (46 per 100). That is, the model is achieving comparable precision despite a significantly lower base rate. The lift from a model predicting this aggregate measure of adverse events is 4.4 times the base rate (or equivalently, 4.4 times higher than flagging at random). We can also consider what share of all officers who went on to experience the outcome in the evaluation period were flagged as high-risk at the end of the pre-period, a measure known as *recall*. In this case the highest-risk 1% flagged by the aggregate outcome model went on to account for 5% of all officers with adverse events during the evaluation period.

In the following section, we will see that more years of experience and longer histories can enable models to predict better, including on the rarer, more serious outcomes that are of particular policy concern.

Table 2: Performance of prediction models on predicting adverse events for all officers, using officers' most recent 5 years of administrative records.

	Top 1%			Top 10%		
	Precision	Lift	Recall	Precision	Lift	Recall
Any citizen complaint	0.69	2.29	0.02	0.55	1.83	0.18
EF complaint with phys. contact	0.02	7.68	0.08	0.01	1.54	0.15
Complaint with suspension	0.02	3.22	0.03	0.01	2.25	0.23
Domestic violence complaint	0.02	9.08	0.09	0.01	3.63	0.36
Aggregate of adverse events	0.35	2.96	0.03	0.27	2.28	0.32

2.4 Assessing Risk of Mid-Career Officers

Once an officer makes it past a department's probationary period, police departments generally must, under the current system, adhere to fairly strict conditions under which an officer can be terminated. The goal of predicting high-risk officers then shifts towards identifying those officers most likely to benefit from some support to mitigate that risk. For example, officers may benefit from a discussion with his or her supervisor, in which citizen complaints on file or previous disciplinary actions are discussed. Some officers may benefit from regular therapy or mental health counseling, or additional training. The appropriate support likely depends on the specifics of the officer's history with the department and may change over time as more interventions are evaluated and found to be successful or unsuccessful at improving an officer's well-being and job performance.

In the previous section, we observed that prediction can be challenging when relying on administrative data collected in officers' first eighteen months on the job. Here, we consider the same analysis, but include data records from all sworn officers rather than those at the start of their career. Predictive factors were included up to 5 years prior to the date of prediction.

Prediction performance using data on all sworn officers is reported in Table 2. While prediction of rarer events was not possible using probationary review data alone, having data over a more extended period of the officer's career substantially improves predictability. For example for the top 1% of the predicted risk distribution, relative to the probationary-period model the lift for mid-career officers increases from 1.6 to 2.3 for predicting any citizen complaint, from 0 to 7.7 for predicting complaints of excessive force with physical contact, from 0 to 3.2 for predicting complaint with a suspension, and from 0 to 9.1 for predicting an off-duty domestic violence or substance abuse incident. (The only measure for which predictive power as measured by lift declines is for our aggregate index of adverse events.) These findings mirror those of Chalfin and Kaplan (2020) which notes that a programmatic intervention that transforms the top 10% of officers, ranked according to risk using data from five years on the job, into a median officer could abate as much as 20% of citizen use

of force complaints over a five-year post-period.

These results make it clear that a longer history of administrative records with the department can improve predictability of risk on a broader range of relevant outcomes. Departments can use statistical models such as those presented here to identify which officers may benefit most from supports and other interventions designed to improve officer decision-making, mental wellness and stability, and performance on the job.

3 State-Level Policy Levers

In the previous section, we consider the prospects for "big data" to be useful in identifying police officers who are at high risk of engaging in out of policy actions which result in harms for the communities they serve. In this section, we describe several ways in which states can assist local agencies in improving their hiring, oversight and support of the police officers they employ. As a general matter there is no guarantee that state involvement in local policy decisions will inevitably improve local outcomes as local residents would define 'improvement.' For that reason we attempt to identify some over-arching principles about why state involvement has some objective reason for potentially improving outcomes, or where only state involvement could accomplish some outcome, namely leveraging economies of scale, solving information failures, and redistributing resources among local areas.

3.1 Solving information externalities

One of the most obvious ways that state policymakers can assist local law enforcement leadership is by solving *information externalities*, a problem which in which organizational decision-making is hampered by incomplete or asymmetric information. We focus on three particular types of information externalities that can potentially be solved with the assistance of state policymakers.

Each year municipal law enforcement agencies terminate thousands of police officers for cause. Yet others quit during the course of an investigation and before they can be officially terminated. Making that information known to other departments to use in making their own hiring decisions generates some cost to the department that fired the officer, but creates a benefit for the other departments to which the former officer might apply. This is what we mean by an externality, and standard textbook economics predicts that departments left on their own will under-invest in making that information widely available.

In the US, a database called the National Decertification Index is used to track individuals for whom a state-level decertification process has decided that the individual should no longer be allowed to exercise the duties of a law enforcement officer. However, a number of states do not engage in decertification or delicensing practices, and the reasons for an officer's release may still be highly relevant for other departments even if it did not result in decertification.[7]

[7]States without official decertification procedures are California, Massachusetts, New Jersey,

Instead, municipal police departments must rely on standard background checks, which do not necessarily uncover behavior that is not of a criminal nature, and on informal reference checks, which may involve co-workers who are hesitant to reveal the nature of an officer's separation.

What becomes of officers who are terminated for cause in the United States? Recent research by Grunwald and Rappaport (2019) shows that a plurality of these "wandering officers" obtain employment with another police agency and, upon being re-hired, are at greater risk for future misconduct. This suggests that states (or possibly even the federal government) can assist municipal leaders by developing and maintaining an active database of police officers who are terminated for cause or who quit during an active misconduct investigation. While local departments would retain the option to hire officers appearing in the database, insurance companies may potentially view such a hire as a liability and could consider raising insurance premiums in response.

Another example of a sort of information externality that states could potentially manage relates to variance in psychological screening procedures for new hires. Job applicants may exploit the variance in psychological evaluation protocols and seek out psychologists who will agree to an applicant's mental and emotional stability (Dantzker, 2011). This will obviously reduce the signal value of such evaluations for identifying future problem officers. States could help prevent this problem by standardizing screening procedures, and creating a database that combines screening information collected by multiple departments.

Finally, local agencies often find themselves competing for the same qualified job applicants. This creates first-mover advantages that could lead to an arms race in hiring that could potentially lead to rushed hiring systems that adversely affect hiring outcomes. States can potentially play a role here in standardizing hiring deadlines so as to avoid this sort of scheduling arms race.

3.2 Redistributing resources

Education, like policing, is a public-sector activity that relies heavily on local funding. This can lead to large disparities in the tax bases, and hence tax revenues, available to support these government functions across localities within a given state. The result can be large inequalities across jurisdictions in the quality of some local public services where for example we see large differences in per-pupil school spending levels between affluent suburbs and nearby city school systems that serve more economically disadvantaged students. In education, we have seen a wave of lawsuits starting in the 1980s in which plaintiffs have sued states for violating constitutional guarantees that children be provided adequate educations, demanding that states increase resource redistribution from rich areas to less affluent localities to support public schooling (Ryan, 2007).

and Rhode Island.

The same resource inequality we see in public schooling is also relevant for public safety. In principle, states could make the policy decision to use the state's ability to raise revenue to effectively redistribute resources from higher income to lower income jurisdictions. More speculatively, there might even also be a legal claim to be made to force this type of redistribution. For example Lowy and Sampson (2016) argue that the Constitution's guarantee of 'life, liberty and pursuit of happiness' could potentially over-ride the Second Amendment's guarantee of the right to keep and bear arms; in principle the argument could be extended to justify redistribution of resources to ensure some level of public safety adequacy in under-resourced communities.

This is relevant because the experience of many previous police departments that have gone through reforms have found that the reform process can be very resource intensive. This certainly seems to be the case, for example, with departments that have successfully completed consent decrees with the US Department of Justice's Civil Rights division. There is no principled reason why the likelihood that one's local police department is able to meet consent decree requirements should depend on the nature of the local property tax base where one lives.

Another local police function that might benefit from state redistribution of resources are the organizations designed to 'police the police.' Beginning with the seminal contributions of Berkeley, CA police commissioner August Vollmer, local leaders have been actively engaged in efforts to reform police departments for more than a century (Douthit, 1975). Like contemporary police reform efforts, early reforms included a robust focus on increasing accountability among rank-and-file police officers. Sixty years ago, Vollmer's student, O.W. Wilson, was among the first police leaders to call for the establishment of a non-partisan police board to help govern the police force upon becoming the police commissioner in Chicago in 1960. Little is currently known about the degree to which these agencies are adequately funded, whatever that might even mean in either principle or practice. But there are certainly indications in the public discussion of widespread dissatisfaction with the work of these agencies, including the long lags often involved in resolving citizen complaints against police, and what are in many jurisdictions fairly low rates of sustaining citizen complaints. Understanding more about the performance of these investigatory agencies, and the potential effects of providing them with additional resources, is one of our recommendations for future research.

3.3 Economies of Scale

Another way states can help localities implement changes to how they manage risk of misconduct is by leveraging opportunities for *economies of scale*. Outside of policing a common example of economies of scale can be found in marketing. While individual McDonald's restaurants would likely find it prohibitively costly to develop a television commercial and pay for it to air, they can gain access to advertising done by the McDonald's Corporation on

behalf of their franchises. Because a single commercial can serve hundreds or even thousands of McDonald's restaurants, there is no need to "re-invent the wheel" by paying to produce thousands of different commercials.

This is not just a useful example to illustrate economies of scale in general, it also suggests a concrete task states can help localities with: advertising to recruit job applicants. The higher quality the applicant pool, the better the subsequent policing outcomes. Indeed a recent field experiment by Linos (2018) suggests that messaging can matter — ads that focus on the personal benefits of police work rather than messages which focus on pro-social motivation are most effective in attracting high-quality candidates to police agencies.

With respect to policing, there are a wide array of other costly tasks — for example, data collection and analysis — which can be done by a single entity on behalf of dozens or hundreds of local law enforcement agencies spreading the fixed costs out so as to maximize productive efficiency. In this section, we propose several concrete ways in which state leaders can take advantage of economies of scale, in particular in the area of data collection and analytic support.

3.3.1 Economies of scale in prediction models

As noted in the previous section, statistical models that predict officer risk of misconduct can be used to help inform police department decisions about hiring, which officers to retain at the end of their probationary periods, and which officers, once hired, would benefit most from supports like training or mental health treatment (as part of a so-called 'early warning system').

One source of economies of scale comes simply with the statistical modeling itself. Developing a credible early warning system requires a great deal of technical expertise which is not typically found among police department personnel. Likewise, hiring civilian employees with the training to develop sophisticated machine learning-based tools is beyond the budget of all but the largest and wealthiest law enforcement agencies, many of which lack even the funds to hire a basic civilian crime analyst. Developing an early warning system has high fixed costs but, critically, those costs do not rise much with the number of officers who require a prediction. So while it would be extraordinarily costly for each police agency in a state to develop its own early warning system on its own, the average cost per police agency from developing a single early warning system at the state-level will be far lower. Naturally, the creation of an early warning system at the state-level cannot ignore the contextual factors that shape police work in different environments. While there is always a tradeoff to aggregating up, there may be technical solutions to this problem including conditioning on an officer's police command and work shift in large agencies, and conditioning on a municipality's crime environment for smaller agencies.

A second and related source of economies of scale is in providing adequate sample sizes to support predictive modeling. While the academic research invariably studies large law

enforcement agencies in big cities, nearly half of police agencies in the United States employ fewer than 10 officers and nearly 80% employ fewer than 25 officers.[8] As discussed above, for the most serious outcomes that are of greatest policy concern there can be too few cases even in large cities to develop predictively accurate models. That problem is greatly exacerbated for smaller jurisdictions. By combining data from hundreds of agencies, the development of an early warning system at the state level has the potential to overcome the enormous challenge of making predictions in small departments which employ few officers and which therefore have little data.

3.3.2 Improving Information Collection and Dissemination

To capitalize on potential economies of scale in the development of prediction models at the state level, it will also be necessary to harmonize the data elements and data collection procedures across local police departments. Recognizing that facilitating cooperation between local agencies is a unique role of higher levels of government, state and federal agencies have taken the lead in creating the infrastructure for local agencies to share information about crime patterns which cross jurisdictional boundaries and suspects with criminal records in multiple states or counties. Today, these initiatives — including the Automated Fingerprint Identification System (AFIS), the National Crime Information Center (NCIC) and DNA databases in all fifty states — form the basis for a variety of core law enforcement tasks including apprehending fugitives and investigating crimes.

Just as state and federal agencies have created the infrastructure for local agencies to cooperate in their mission to maintain law and order and enforce the law, they can play a similar role in assisting local agencies with managing their personnel. By recommending — or requiring — that local agencies collect and share a standardized set of information about job applicants, new hires and officers on the job, states can facilitate the creation of predictive models that will service all of its constituent local agencies and can continue to develop data-driven insights that make policing safer for both officers and the communities they serve.

One particularly useful form of data collection that states might focus on is collection directly from the public itself of additional measures of officer performance, to use as additional dependent variables in prediction models. Publicly-reported statistics show that in Chicago, for instance, citizen complaints accrue at a rate of 0.3 per officer per year (Chalfin and Kaplan, 2020). The protests we have seen across the country in 2020 show that, in the public's mind, figures like this are clearly too high. But from a statistical modeling perspective, the relatively low density of complaint data statistically speaking, serves as a natural constraint on the ability of an early warning system or other predictive model to identify the highest-risk officers.

[8] See: https://www.bjs.gov/content/pub/pdf/lpd13ppp.pdf

How can complaint and related administrative data be supplemented in order to capture other potentially useful dimensions of on-the-job performance? Since officers operate independently and are rarely visible to their supervisors, measuring officer performance is an extraordinarily difficult task. However, this feature of police work is far from unique. Typically, in industries in which employee activities are difficult for supervisors to monitor, employers lean heavily on data obtained directly from customers. Soliciting customer feedback is a mainstay of many industries and allows employers to develop a richer understanding of the extent to which employees provide consistent and high-quality service to customers.

Customer service is an important part of a police officer's job. Research using the Police Public Contact Survey, a nationally representative survey on police service, consistently shows that individuals who perceive that they received poor police service have more negative views of law enforcement are are less likely to report crimes in the future (Slocum, 2018). While most salient to the public are probably police contacts with criminal suspects, in fact the majority of contacts that members of the public have with police arise due to car accidents or a contact with a citizen who was the victim of or witness to a crime. More than 50 million Americans — or 21% of the population over the age of 16 — report contact with a police officer each year in the United States. Many individuals report more than one contact. So police contacts represent potentially fertile ground to learn about police behavior. 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 turn out to contain useful signal for identifying high-risk officers. And for the reasons described above, state involvement may be necessary to ensure standardization in what is collected so that this information can be put to socially productive use.

3.3.3 Improving Supports to Reduce Risk of Misconduct

A final source of economies of scale that states can help capitalize on has to do with the provision of training and mental health supports to officers who are at high risk for adverse outcomes. When officers are flagged by an early warning system, the usual desire is to match them to effective supports as quickly as possible to mitigate that risk. But because of the fixed costs associated with running a training, small departments may provide a given type of training only infrequently. Pooling efforts across local departments as the state could help do would make it easier for every department to ensure there are immediately available (and ideally high quality) supports available for officers when these supports are needed.

Similarly state involvement can help deal with economies of scale in the evaluation of these supports. In the area of crime prevention, decades of research show us that most prevention programs don't work. The lesson seems to be that behavior change is typically difficult to achieve. There is no reason to believe this is intrinsically any easier when program participants are police officers rather than civilians. So rigorous evaluation of risk-mitigation

supports is likely to be an important part of any serious effort to reduce prevalence of police misconduct in the United States. For any individual department, there may be too few officers receiving a given training over any defined time period to support a rigorous evaluation. But if trainings or mental health supports are standardized and pooled across departments, states would be helping to ensure that there are enough people participating in a given defined type of support to allow for a rigorous assessment of its causal impacts on risk of adverse outcomes.

4 Future Research

We conclude by suggesting several avenues for future research. First, due to constraints on the availability of police officer microdata, there are only a handful of papers that delineate the predictors of police misconduct. These papers cover only a small number of large law enforcement agencies and often use only a subset of the data that are potentially available to make predictions. As a result, it is difficult, at present, to have a detailed understanding of how easy it is to predict high-risk officers in different types of police departments and at different stages of their career. Collection of new types of data at the application and probationary period stage could potentially improve the performance of prediction models at these early career points, including potentially new types of measures informed by behavioral science.[9] Additional work is needed to understand how to tailor assessment tools and identify indicators that are most predictive of later success as a police officer. In our view, the development of this research area is an urgent priority.

Second, even if misconduct is highly predictable, it is unclear how departments can best intervene in order to mitigate the risks posed by high-risk officers. While departments can decline to hire prospective officers at the time of their employment application and can terminate high-risk officers during the standard probationary period, misconduct is, in general, more difficult to predict at these stages than it is later in an officer's career. Making hiring requirements more stringent could have implications for the overall pool of potential applicants, akin to those observed from increasing certification requirements for teachers (Angrist and Guryan, 2008). Given that it is difficult to terminate mid-career officers especially when the basis for termination is a prediction, law enforcement agencies would benefit from better understanding of effective ways to re-train or otherwise support high-risk personnel in order to mitigate that risk. 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)

[9]For example in the private sector, the start-up Pymetrics has combined behavioral science approaches to baseline application measurement with algorithms that use these measures to predict on-the-job performance to apparently help both increase hiring diversity and potentially improve productivity as firms define it.

and procedural justice training (Owens et al., 2018; Nagin and Telep, 2020; Wood et al., 2020) as well as the use of and training in non-lethal weapons (MacDonald et al., 2009; Sousa et al., 2010). However, these interventions have been broad-based measures to train a representative group of officers and none of these interventions are designed specifically to remediate personnel once they have been administratively designated to be high-risk.

Across a wide range of policy policy domains we are seeing a movement towards 'evidence-based policy,' that uses data to both create new interventions and evaluate their effectiveness. Criminal justice is no exception. We have outlined here a number of ways in which data could potentially be helpful for the effort of reducing police misconduct. States, by virtue of their unique ability to coordinate across agencies and pool or redistribute resources, may have several important roles to play in such efforts.

References

Aaronson, D., L. Barrow, and W. Sander (2007). Teachers and student achievement in the chicago public schools. *Journal of Labor Economics 25*(1), 95–135.

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

Angrist, J. D. and J. Guryan (2008). Does teacher testing raise teacher quality? evidence from state certification requirements. *Economics of Education Review 27*(5), 483–503.

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

Ater, I., Y. Givati, and O. Rigbi (2014). Organizational structure, police activity and crime. *Journal of Public Economics 115*, 62–71.

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

Bissett, D., J. Bissett, and C. Snell (2012). Physical agility tests and fitness standards: perceptions of law enforcement officers. *Police Practice and Research 13*(3), 208–223.

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

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. *American Economic Review: Papers and Proceedings 106*(5), 124–127.

Chalfin, A. and F. Goncalves (2020). The pro-social motivations of police officers. *Working Paper*.

Chalfin, A. and J. Kaplan (2020). How many complaints against police officers can be abated by incapacitating a few 'bad apples?'. *Available at SSRN 3673981*.

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*.

Dantzker, M. (2011). Psychological preemployment screening for police candidates:seeking consistency if not standardization. *Professional Psychology: Research and Practice 42*(3), 276–283.

Douthit, N. (1975). August vollmer, berkeley's first chief of police, and the emergence of police professionalism. *California Historical Quarterly 54*(2), 101–124.

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*.

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

Grunwald, B. and J. Rappaport (2019). The wandering officer. *Yale Law Journal 129*, 1676.

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

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*.

Linos, E. (2018). More than public service: A field experiment on job advertisements and diversity in the police. *Journal of Public Administration Research and Theory 28*(1), 67–85.

Lonsway, K. (2003). Tearing down the wall: Problems with consistency, validity, and adverse impact of physical agility testing in police selection. *Police Quarterly 6*(3), 237–277.

Lowy, J. and K. Sampson (2016). The right not to be shot: Public safety, private guns, and the constellation of constitutional liberties. *The Georgetown Journal of Law Public Policy 14*(1).

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.

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

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.

Persico, N. and P. Todd (2006). Generalising the hit rates test for racial bias in law enforcement, with an application to vehicle searches in wichita. *The Economic Journal 116*(515), F351–F367.

Roberts, R. M., A. M. Tarescavage, Y. S. Ben-Porath, and M. D. Roberts (2019). Predicting postprobationary job performance of police officers using cpi and mmpi–2–rf test data obtained during preemployment psychological screening. *Journal of Personality Assessment 101*(5), 544–555.

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.

Ryan, J. E. (2007). Standards, testing, and school finance litigation. *Texas Law Review 86*, 1223.

Sanders, B. A. (2008). Using personality traits to predict police officer performance. *Policing: International Journal Police Strategies & Management 31*, 129.

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

Slocum, L. A. (2018). The effect of prior police contact on victimization reporting: Results from the police–public contact and national crime victimization surveys. *Journal of Quantitative Criminology 34*(2), 535–589.

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.

Staiger, D. O. and J. E. Rockoff (2010). Searching for effective teachers with imperfect information. *Journal of Economic Perspectives 24*(3), 97–118.

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.

Weisburst, E. (2020). Whose help is on the way? the importance of individual police officers in law enforcement outcomes. *Working Paper*.

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, X. and C. Lum (2017). Measuring the spatial and temporal patterns of police proactivity. *Journal of Quantitative Criminology 33*(4), 915–934.

Zhao, J. M. C. Scheider, and Q. Thurman (2003). A national evaluation of the effect of cops grants on police productivity (arrests) 1995-1999. *Police Quarterly 6*(4), 387–409.