Metrics Management and Bureaucratic Accountability: Evidence from Policing [*†]

Laurel Eckhouse [‡]

November 24, 2020

[*] Many thanks to Amy Lerman, Gabriel Lenz, Sean Gailmard, Hannah Walker, Ariel White, Allison Harris, Amelia Hoover Green, Sherry Zaks, Rachel Bernhard, and three anonymous reviewers for extensive and helpful comments. I am also grateful to Ruth Collier, Courtenay Conrad, Adrienne Hosek, and participants in the Visions in Methodology Conference, as well as to the Research Workshop in American Politics at the University of California, Berkeley, for valuable suggestions and feedback.. For research assistance, I thank Ada Lin, Benjamin Huber, Nicole Restmeyer, and Shireen Ebrahim. For data sharing, I thank Justin de Benedictis-Kessner. This research was supported by funding from the National Science Foundation Graduate Research Fellowship Program, the Horowitz Foundation for Social Policy, and the Travers Department of Political Science at UC Berkeley.

[†] Keywords: policing, bureaucracy, management, metrics management, criminal justice, police, law enforcement

[‡] Assistant Professor, Department of Political Science, University of Denver. laurel.eckhouse@du.edu

Abstract

Bureaucracies increasingly use quantitative measures to monitor personnel behavior. I develop a model of the incentives created by *metrics management*, a bureaucratic accountability technique, using policing as a case to show that monitoring can lead public-interest motivated bureaucrats to focus on work not in the public interest. Second, I develop a new measure of data manipulation in crime statistics: while theory predicts the presence of manipulation, researchers observe only the altered data. I solve this using the fact that police departments can reclassify rapes (but not other violent crimes) as “unfounded”, concluding the reported crime did not occur. Finally, I test the effects of metrics management in policing using a novel data set. Compstat is associated with at least 3,500 additional minor arrests per city-year, substantial data manipulation, and no decrease in serious crime. These results have implications for bureaucracies implementing metrics management, scholarship using administrative data, and legal implementation.

Word count: 9,999 words

The data and materials required to verify the computational reproducibility of the results, procedures and analyses in this article are available on the American Journal of Political Science Dataverse within the Harvard Dataverse Network, at:
https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/3E7JXB

1 Introduction

Government services depend on "street-level bureaucrats": agents who make and enforce bureaucratic decisions (Lipsky, 2010). Legislative principals choose policies, but agents implement them, often with minimal supervision and tremendous discretion. Small wonder, then, that bureaucratic politics has become a central site for the study of principal-agent delegation (Meier and Krause, 2003).

The history of policing exemplifies the principal-agent problems of the bureaucracy. Police have a long history of shirking, including sleeping on duty – a practice so common it has its own name and a century history of managerial intervention (Zacks, 2012). When awake, police work unobserved in high-stakes, ambiguous situations. Practically speaking, "discretion is inevitable" (Wilson, 1978): police will never be able to enforce every law at every moment (Stuntz, 2011).

In the early 1990s, police departments developed new tools for controlling bureaucrats. Drawing on technological improvements and the New Public Management, police departments monitored crime and arrest statistics, then held officers and supervisors accountable. They focused on quantitative indicators, developing a system I call *metrics management*. Metrics management spread far beyond policing, promising to stop bureaucratic shirking and hold agents accountable for results. While measuring street stops and drug arrests is easy, measuring community engagement and crime prevention is hard. Metrics management encourages bureaucrats to prioritize work with rapid, measurable results – and deprioritize complex, uncertain, potentially more important projects.

This article makes three contributions to the bureaucratic accountability literature. First, I develop a formal model of metrics management, using policing as a case study to show how measurement changes bureaucratic incentives. These results expand our understanding of agency models, showing how, when success is uncertain, monitoring can lead bureaucrats to work that does not serve the public interest – even when they share their principals' motivations.

Second, I develop a new measure of data manipulation in crime statistics. Measuring the manipulation predicted by the model is challenging, since only the altered data are recorded. To solve this problem, I draw on the ability of police departments to reclassify rapes (but not other violent crimes) as "unfounded" by concluding the reported crime did not occur – a type of manipulation that leaves traces in national datasets.

Finally, I test the model's implications using a novel data set on metrics management in police departments. I find increased minor arrests and data manipulation. Minor arrests increase by at least 3,500 per city-year, a large substantive change with major feedback consequences. Metrics management is associated with a substantial increase in the designation of rapes as unfounded, suggesting that crime clearances[1] and statistics are likely skewed by data manipulation under metrics management.

I use policing as a case study because the stakes are high: the consequences of shirking, which metrics management aims to prevent, are serious, as are the consequences of increases in minor arrests and data manipulation. People's lives and civic participation may be forever altered by an arrest (White, 2019; Kohler-Hausmann, 2013; Walker, 2018), or by the lack of state consequences for serious crimes when they are downgraded or hidden (Leovy, 2015).

The model of incentives suggests that the consequences – reduced shirking, at the cost of reductions in complex projects and data quality – should be similar in other bureaucratic domains that adopt metrics management. Metrics management has spread through schools, wartime body counts, the nonprofit sector, even to legislative report cards. These numbers are politically and socially consequential: mayors, school superintendents, and national politicians tout "good numbers" in campaigns, schools close when they fail to make Adequate Yearly Progress, and teachers, police officers, and other bureaucrats reshape their jobs around metrics.

Finally, I check for changes in the incidence of crime, and in crime clearance rates – outcomes anticipated by proponents of metrics management. I find no change in the incidence of violent crime. Clearance rates do improve. However, evidence of data manipulation, and examples from journalists and ethnographers of untraceable manipulation, cast doubt on this improvement.

[1] A crime is "cleared" when a suspect is arrested, or by "exceptional" means, e.g. police identify a perpetrator who is dead or cannot be extradited, or conclude the crime did not occur. (Federal Bureau of Investigation Staff, 2012)

1.1 The Stakes: Metrics Management in Policing

The stakes of metrics management are particularly high in policing. Where officers are accountable for quantitative targets, they are often expected to demonstrate productivity via minimum numbers of arrests, citations, or other interventions, and face career consequences if they fail to meet these targets (Jeffers, 2014; Moskos, 2009; White, 2008). Skyrocketing criminal legal contact in the 1990s, with well-documented negative consequences, was due in part to changes in police incentives (Brayne, 2014; White, 2019; Goffman, 2014). The thousands of additional minor arrests per year associated with metrics management expose thousands of civilians to these consequences.

Data manipulation also has massive substantive consequences. Media accounts of the New York, Chicago, and Los Angeles police departments have described police downgrading of serious crimes. In New York, a string of related sexual assaults occurred in a single neighborhood in 2002. Under pressure to post weekly crime reductions, police recorded the assaults as criminal trespassing or other misdemeanors. Over two months, the perpetrator grew bolder in his methods and targets. The pattern of attacks was not discovered until he was apprehended and confessed, when the detective questioning him looked through the precinct's complaints and found the misclassified incidents (Rayman, 2013).

This case exemplifies the real-world consequences of data manipulation by police. Communities are robbed of public safety and access to formal law when officers downgrade complaints. In Chicago, police manipulated the number of robberies and, shockingly, the number of homicides, typically considered the most challenging statistic to answer. Police hid homicides by reclassifying suspicious deaths as "noncriminal death investigations" to avoid increasing the city's murder rate. These noncriminal deaths included one where a woman's naked body was found in an abandoned warehouse, with evidence that she had been tied to a chair and gagged (Bernstein and Isackson, 2014). A Los Angeles Times investigation found the LAPD misclassified nearly 1200 violent crimes as minor incidents (Poston and Rubin, 2014). My national analysis of data manipulation suggests that this is a systemic outcome of metrics management.

2 Metrics Management in Bureaucratic Context

In the 1980s and 1990s, the New Public Management encouraged public agencies, including police, to decentralize accountability and develop incentives for performance (Andrisani, Hakim and Savas, 2002; McLaughlin, Osborne and Ferlie, 2002; Gruening, 2001). Combined with technological improvements, police departments developed – and other agencies soon adopted – metrics management. In 1994, New York Police Commissioner Bill Bratton pioneered Compstat, the first such system for policing. Compstat's technological component allowed police managers to visually display crime data from recent events. The NYPD used Compstat data to find patterns, set targets, and evaluate the performance of precinct commanders and street-level officers. Meetings were frequently combative: precinct commanders and superiors were responsible for specific geographic areas, and were expected to report successful results at meetings with top police and mayoral officials (Eterno and Silverman, 2012; Henry, 2002). These meetings expanded monitoring throughout the chain of command: mayors monitored the police chief, the police chief monitored area commanders, area commanders monitored front-line staff. Similar systems – CitiStat, ParkStat, ATLstat, and SFstat – spread to education, city administration, and other bureaucracies (Eterno and Silverman, 2012).

Metrics management has appealing promises. It reduces information asymmetry between policymakers and bureaucratic agents (Miller, 2005), aligns the incentives of street-level bureaucrats with those of policymakers, and lets departments efficiently oversee agents' performance (McCubbins and Schwartz, 1984; McCubbins, Noll and Weingast, 1987). Moreover, it addresses a core problem for bureaucracies: with discretion comes the risk of shirking (Brehm and Gates, 1993).

Shirking dominates the academic literature on police bureaucracies (Brehm and Gates, 1997; Allen, 1982; Engel, 2000). It is also a core concern for police managers. In the New York Police Department, "cooping" – slang for sleeping on duty – has been a concern since Theodore Roosevelt rousted sleeping officers in the early 1900s (Zacks, 2012). More recently, it was so widespread that the NYPD banned officers from the best nap spots (Goldstein, 2014). When officers are awake, they work unobserved, making it hard to control their actions. Body-worn cameras have created the possibility of observing officer activity, but selectively and after the fact, using “fire alarms” rather than “police patrols” (McCubbins and Schwartz, 1984). Similar concerns apply in other bureaucracies: will teachers work hard once the classroom door is closed? Will regulators enforce food safety rules or environmental protections unsupervised?

Measurement has pitfalls, though. Journalists have caught police departments reclassifying homicides, downgrading rapes to less serious crimes, and misclassifying assaults (Rayman, 2013; Bernstein and Isackson, 2014; Poston and Rubin, 2014). Schools under pressure to meet test score targets have seen cheating scandals (Vogell, 2011; Aviv, 2014). Data manipulation is not the only concern about quantitative targets: in places where Compstat and Broken Windows policing have been implemented, officers report pressure to meet quotas for arrests and tickets. Because departments track these productivity measures, officers express concern that if their arrest rates are low, they will face career consequences (Jeffers, 2014; Moskos, 2009; White, 2008).

Scholars of bureaucratic politics have noted the pathologies of measurement at least since Blau found that monitoring employment counselors’ placement rates displaced their focus towards easier clients (Blau, 1963; Wilson, 1989). Since then, many agency models have engaged the problems of monitoring and supervision for bureaucratic agents (Brehm and Gates, 1993; Gailmard and Patty, 2012, 2007; Holmstrom, 1982; Tsebelis, 1989; Bianco, Ordeshook and Tsebelis, 1990). Still, metrics management represents an important innovation which is not well described in existing qualitative research on agency loss and goal displacement.

Measurement changes behavior in two ways. First, it leads bureaucrats to pursue demonstrable productivity, like minor arrests, even when this does not contribute to larger goals like reducing crime. Even when officers would rather focus on longterm projects, monitoring can lead to a focus on short-term work with reliable, legible results. Second, measurement can lead to data manipulation: officers or other bureaucrats may change recording or reporting practices to make their observable results look better, without changing their behavior or the desired outcomes.

These two pathologies of measurement are manifestations of the same problem. Incentive changes designed to reduce shirking change the mix of tasks that street-level bureaucrats pursue, increasing incentives for police to pursue strategies with guaranteed outcomes, like minor arrests and data manipulation, rather than invest in more challenging, less certain projects. Supervisors face inherent challenges – not just resource constraints – in discouraging shirking while encouraging longterm probabilistic work like building community relationships and investigating serious crimes (Goldstein, Sances and You, 2018). The formal model identifies these tradeoffs, then uses them to predict the effect of monitoring on bureaucratic compliance.

Metrics management differs in two ways from previously identified examples of goal displacement from monitoring. First, with better computational tools and quantitative analysis, supervisors combined the historical focus on work activity with measures of actual performance (Lipsky, 2010). Metrics management differs from older techniques in also holding agents accountable for the outcomes of bureaucratic work: test scores, crime rates, street conditions, and highway fatalities. Outcomes are partially outside bureaucrats' control, creating stronger incentives for data manipulation. However, work activity, like citations and arrests, still plays a role in metrics management.

Second, previous literature on street-level bureaucrats emphasizes horizontal accountability and the influence of street-level bureaucrats (Bovens, Goodin and Schillemans, 2014). Historically, bureaucrats have shaped the measures of their performance, protecting themselves from intrusive measurement techniques and maintaining independence from supervisors (Lipsky, 2010). Metrics management, in contrast, imposes supervision from above to generate data supporting principals' electoral or market incentives: scores, crime rates, or quarterly earnings measures that shareholders or voters will reward. These often require intrusive data collection that agents object to, like high-stakes testing and elaborate documentation (Mummolo, 2015).

While Kelley and Simmons (2015) find measurement and quantification per se exert social pressure on bureaucrats, metrics managers also reward (and punish) bureaucrats for results. Thus, metrics management provides a valuable context for examining problems with bureaucratic monitoring, discretion, and goal displacement (Langbein, 2010; Andersen and Moynihan, 2016; Meier, Polinard and Wrinkle, 2000). Formal models of agency and delegation are particularly suited to analyzing these dynamics. Below, I describe the work typologies for bureaucratic decisions under metrics management. Then, I present a formal model of metrics management, and explain its contributions to the literature on agency models.

2.1 Monitoring Probabilistic Work

Metrics managers face a fundamental challenge: what should they monitor? Types of bureaucratic work fall into two primary categories: *longterm* work, which is probabilistic in nature and does not deliver immediately observable results; and *guaranteed* work, which has immediate outcomes. I formalize this decision problem as a two-player game with an Agent and a Supervisor. The key intuition is that shirking is observationally equivalent to a longterm project that has not yet yielded results. Increases in monitoring reduce shirking, but also reduce the payoff for longterm work, shifting the Agent's work to short-term projects with guaranteed results.

In conversations with journalists and ethnographers, officers describe many types of *longterm* work: building community relationships, investigating complex cases, cultivating informants, or mediating disputes (Moskos, 2009; Rayman, 2013). These complex projects often have few observable results. Until these projects come to fruition – when charges are filed against the criminal network, or community relationships yield leads after a crime – working on them is observationally indistinguishable from shirking. Moreover, *longterm* projects are inevitably probabilistic. A competent investigation could fall apart at the last minute; a community relationship might be interrupted when a civilian leaves the neighborhood, or an officer is redeployed. This makes valid measurement of *longterm* efforts challenging.

However, bureaucrats can also typically pursue *guaranteed* work: tasks that can be reliably completed, with clear, demonstrable results. Enough Americans use and carry drugs for officers to reach an arbitrary number of drug arrests by expanding stops and searches (Stuntz, 2011). Detectives may prioritize straightforward cases with known perpetrators, make arrests despite missing evidence so they can designate a case cleared, claim they have identified a perpetrator who cannot be arrested (clearing the case by extraordinary means), or reclassify the crime as a lesser offense

(Rayman, 2013; Thompson, 2000; White, 2008).

Data manipulation is a form of *guaranteed* work that manipulates outcomes rather than work effort: it has a reliable payoff, which officers can be certain of achieving on a clear time-scale. While departments and individuals sometimes face backlash for data manipulation, it typically comes long after the accountability portion of Compstat is complete, and only via large-scale journalistic or governmental investigations (Rayman, 2013). On the timescale of daily decision-making, data manipulation is reliable for officers. (For more on manipulation versus guaranteed work, see Appendix E, p. 16).

The same strategies are available in other bureaucracies. Teachers, working unsupervised, may $shirk$, minimizing their time planning lessons and offering student feedback. They may spend classroom time on *longterm*, probabilistic work: supporting independent inquiry, self-monitoring, and complex academic skills. Or, they may focus on *guaranteed* strategies for improving test scores, substituting test prep for other classroom activities (Abrams, Pedulla and Madaus, 2003). Moreover, current evidence suggests that the rise of high-stakes testing has fostered data manipulation: teachers and principals reviewing tests in advance, falsely designating students as learning disabled to get them extra time, even changing student answers to ensure passing scores (Vogell, 2011; Aviv, 2014).

When bureaucracies adopt metrics management, the probability of monitoring rises, so the cost of shirking or apparent shirking increases. Since shirking and *longterm* work may have identical results from a manager's perspective, agents shift away from *longterm* work, towards various types of *guaranteed* work. In contrast to models focused on the match between the preferences of the agent and the principals (Brehm and Gates, 1997; Gailmard and Patty, 2012), under metrics management, even bureaucrats who prefer *longterm* projects will shift to *guaranteed* work, unless their preference for *longterm* projects is very strong.

2.2 Formal Model

The two players, an Agent and a Supervisor, correspond to a police officer and a superior, like a sergeant or precinct commander. The dynamics apply to other bureaucratic situations: lieutenants and police chiefs, police chiefs and mayors, principals and teachers, regulators and agency heads.

At the outset, the Supervisor sets the probability of monitoring, m, known to both Supervisor and Agent. Monitoring lets the Supervisor see the Agent's work before payoffs are distributed, rather than after. The Agent then chooses to $shirk$, pursue $guaranteed$ work, or invest in a $longterm$ project. Choosing a $longterm$ project has an immediate cost e in effort. The project will $succeed$ with probability p, yielding a benefit, and otherwise $stall$, leaving the Agent no observable accomplishment. The overall payoff to choosing $longterm$ in the absence of monitoring is y_l, including a fixed effort and a probabilistic reward (that is, $y_l = pr - e$, which I collapse for simplicity).

Choosing $guaranteed$ work yields a payoff y_g, net the cost of working. Choosing $shirk$ has no direct cost to the Agent, and yields a benefit of y_s. After the Agent chooses a strategy, the Supervisor monitors with the probability chosen at the outset. If the Agent has chosen $shirk$, or if $longterm$ work has $stall$ed, the Supervisor imposes a penalty z; otherwise, the Agent keeps the payoff. Preferences are exogenous to the game; the Agent prefers to $shirk$ in the absence of monitoring, but may prefer either $guaranteed$ or $longterm$ work. The Supervisor receives some benefit $a > 0$ if the Agent pursues a successful strategy, and nothing if the Agent $stall$s or $shirk$s.

The sequence of play and the payoffs are summarized in Figure 1 and Table 1. The Supervisor's payoffs are such that $pa_l > a_g > 0$: that is, the Supervisor prefers $longterm$ work, but prefers $guaranteed$ to $shirk$.[2]

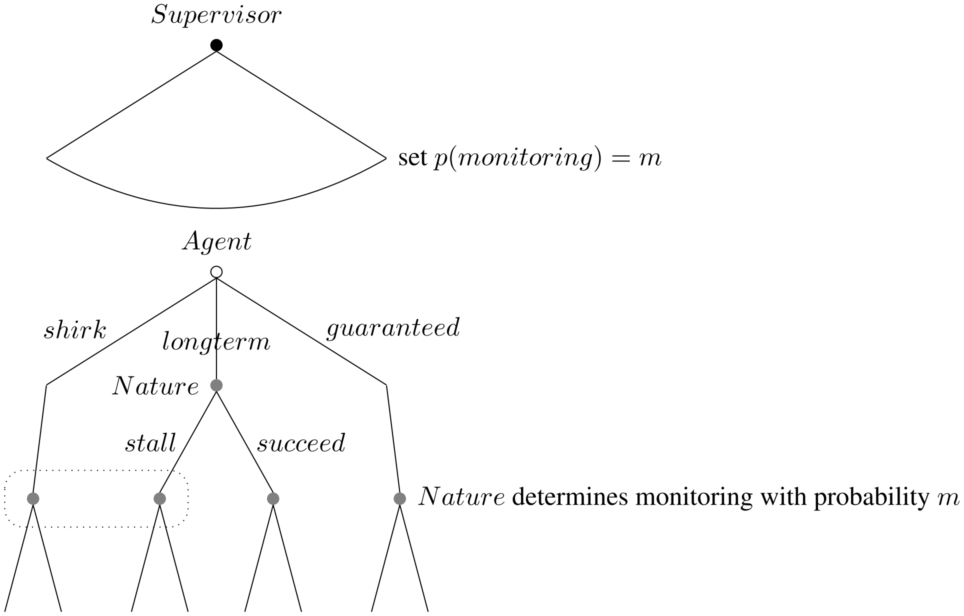
Table 1: Payoffs ($Agent$, $Supervisor$)

$Agent$ strategy	$shirk$	$guaranteed$	$longterm$
payoff	$(y_s - mz, 0)$	(y_g, a_g)	$(y_l - (1-p)mz, a_l p)$

[2] This is the only scenario for the Supervisor's payoffs I consider. If the Supervisor prefers the Agent to $shirk$, the Supervisor need only set $m = 0$; if the Supervisor prefers $guaranteed$, a higher probability of monitoring suffices.

If the Agent preferred $longterm$ to $shirk$, the Supervisor could set $m = 0$. Observation suggests this is an unusual preference among Agents across most domains; I omit this in the model but discuss under Significance.

Figure 1: Game



How can the Supervisor induce the Agent to work? A higher probability of monitoring, m, is formally equivalent to a more severe punishment for choosing $shirk$. If the Agent prefers $shirk$ing to either type of work, the Supervisor must set $m > 0$ such that

$$y_g > y_s - mz \text{ or } y_l - (1-p)mz > y_s - mz$$

or, equivalently,

$$mz > y_s - y_g \text{ or } pmz > y_s - y_l$$

As mz rises the value of $longterm$ work relative to $guaranteed$ declines. The payoff for $guaranteed$ work is not changed by monitoring, but the payoff for $longterm$ work declines by $(1-p)z$ multiplied by the change in the probability of monitoring. The payoff for $shirk$ declines even more steeply, by mz. Unless Agents have very strong exogenous preferences for $longterm$ work – that is,

$$(1-p)mz < y_l - y_g$$

– increasing monitoring will not only decrease shirking, but decrease $longterm$ work.

An Agent who is indifferent between $guaranteed$ and $longterm$ work without monitoring (that is, for whom $y_g = y_l$) prefers $guaranteed$ in the presence of monitoring. Unless the Agent prefers working to shirking, the Supervisor must set $m > 0$ to induce the Agent to choose $guaranteed$ or $longterm$. Unless the payoff of $longterm$ work is already higher than the payoff for $guaranteed$ work by at least $(1 - p)mz$, the Agent will choose the $guaranteed$ option.

This model is agnostic as to the Agent's preferences for $longterm$ versus $guaranteed$ work without monitoring. Police departments, like other bureaucracies, employ individuals with varying values and costs for different kinds of work, depending on their personalities, career goals, and philosophical beliefs. Some agents choose bureaucratic employment precisely to promote the agency's longterm goals (Gailmard and Patty, 2007). Unless Agents are intrinsically motivated not to $shirk$, Supervisors face an inevitable conflict between deterring $shirk$ing and promoting probabilistic $longterm$ work. Moreover, the $succeed$ condition may not be reached by projects which are in other terms "successful": excellent work building community relationships may nevertheless yield no observable results unless and until community support solves a crime.

Observable Implications

The model provides two core observable implications. For a longer version, including proofs, see Appendix E (p. 11).

- If monitoring increases, shirking should decline as the value of $shirk$ falls. The payoff for $longterm$ work also falls as monitoring increases, but more slowly.

- If monitoring increases, $guaranteed$ work increases. As monitoring increases, the expected utility of $guaranteed$ work for the Agent remains constant. Thus, for individual Agents who prefer $shirk$ to $guaranteed$ work by an amount $\leq z$, the payoff from $guaranteed$ work will eventually be greater than the payoff from $shirk$.

  The payoff for $longterm$ also falls with increased monitoring. Since payoffs for $longterm$ and $shirk$ fall with increased monitoring, more monitoring should lead more agents to pursue $guaranteed$, implying an observable increase in routine arrests and data manipulation. This displaces both shirking and other forms of work, suggesting more $guaranteed$ in both absolute and relative terms.

2.3 Metrics Management and Agency Models

This model makes three contributions to the literature on delegation. First, I offer a framework for analyzing monitoring of probabilistic work. The fact that $longterm$ projects and $shirk$ing can be observationally equivalent creates problems for bureaucratic compliance, as the empirical tests show.

Second, I link literatures on agents' preferences and supervision. Brehm and Gates (1993) argue that defection "varies by the attitudes of the subordinates toward policy" , while in Holmstrom's model supervision is the key element (Holmstrom, 1982). Here, supervision shifts incentives regardless of attitudes. Even when bureaucrats are motivated by the public interest, quantitative monitoring can shift their work to activities that do not serve that interest unless their preferences are very strong. The model thus expands the literature on the counter-intuitive effects of monitoring on behavior (Turner, 2017; Carrigan, 2018).

Finally, this model provides tools for assessing different types of agent compliance, and qualitatively different forms of production. Models of bureaucratic effort have typically focused on levels of production. Brehm and Gates (1997) describe the effects of supervision and agency goals on working, shirking, and sabotage – positive, zero, and negative production – and conceptualize sabotage as a way to dissent-shirk when agents disagree with bureaucratic goals, while earlier work contrasts "donut shops" (shirking) with "speed traps" (diligence) (Brehm and Gates, 1993). In contrast, this model distinguishes among qualitatively different forms of compliance.

3 Compstat and Its Consequences

3.1 Compstat: Metrics Management for Policing

Compstat provides a case for studying the consequences of metrics management in bureaucratic practice. Since Bratton and the NYPD developed it, Compstat has become the professional norm among police departments. Departments explain they adopt it with the goal of reducing crime and improving control over field operations. These programs have key shared features: they incorporate specific, measurable objectives; mapped data collection; shifts in both authority and responsibility for data to the managers of geographic subunits; and regular meetings using data to evaluate data (Willis, Mastrofski and Weisburd, 2007). These meetings are meant to induce managers to reduce the incidence of major crimes, but metrics management in policing also involves monitoring data about the performance of individual officers.

Compstat is thus an ideal test case for this model. It is a form of increased monitoring, driven by the twin goals of reducing serious crime and increasing operational control (including less shirking). It has been widely adopted in the United States, but temporal variation in diffusion means that the effects of adoption can be distinguished from overall time trends. National data on crime and arrests allows robust evaluation.

Moreover, policing is a substantively important area of bureaucratic intervention in citizens' lives. This paper sheds light on the major variation in policing regimes across cities. Existing research frames police decisions as the result of the attributes of individual police officers, or of large-scale changes in state and federal criminal law (Weaver, 2012; Murakawa, 2014; Glaser, 2015; Eberhardt et al., 2004). Yet, in 2010, Detroit saw 38.5 arrests per 1000 residents, less than a quarter of Baltimore's 169 arrests per 1000 residents. Tucson had over twice as many arrests per capita as Phoenix. Local bureaucratic politics are key to this disparity: adopting metrics management changes the use of police discretion and increases the use of minor arrests.

Previous evaluations of the consequences of metrics management have focused on individual cities. Reporters and ethnographers have described how departments supervise police officers' day to day activities and the collection of crime statistics. However, this is the first nationwide evaluation of Compstat's effects on police work.

3.2 Observable Implications: Shift from Longterm to Guaranteed Work

In policing, *guaranteed* work takes two observable forms: minor arrests and data manipulation. Arrests necessarily imply a lack of shirking. Recent ethnographers of Black neighborhoods write that "times have changed" since the 1980s, when police left these neighborhoods to their own devices (Anderson, 1978). Now, warrant sweeps and street stops are commonplace. The risk of arrest for minor offenses is high in many cities, especially for young Black and Latino men (Goffman, 2014; Gelman, Fagan and Kiss, 2007; Rios, 2011).

For police officers, investing in major arrests, mediating disputes, and cultivating informants is risky: these activities require substantial resources without guaranteed payoff. Preventive policing and activities that build community relationships are even harder to reward, since they leave no indicator of officers' activity (Moskos, 2009).

The (inevitably discretionary) enforcement of minor crimes, meanwhile, can be reliably performed and yields guaranteed results. Warrant sweeps and arrests for casual drug use are easily performed and documented. As the formal model shows, metrics management pushes police to increase their visible activity in both absolute (reduced shirking) and relative (shift from existing work) terms. Both the number of arrests for minor, consensual crimes, as well as the share of total arrests that are for minor crimes, should increase. Rayman (2013) describes a vivid example of this during the NYPD tapes scandal, in which a rookie officer taped meetings where patrol officers were given arrest and citation quotas.

3.3 A New Technique for Measuring Data Manipulation

When people's careers and livelihoods are affected by data they produce, but the underlying activity is difficult to change, they seek *guaranteed* techniques to reach the numbers their supervisors demand. Reducing serious crime is challenging, giving police reason to manipulate the data by changing reporting patterns. In policing, major crimes may be underreported, because supervisors are rewarded for lowering serious crime; meanwhile, police are expected to demonstrate productivity through specific, quantifiable actions, leading them to overenforce and thus overreport minor violations.

Data manipulation has been identified locally in a variety of crimes: homicides, assault, and sexual assault (Rayman, 2013; Bernstein and Isackson, 2014; Poston and Rubin, 2014). Measuring manipulation is challenging, since researchers have access only to the altered data. However, I identify a strategy for identifying data manipulation in the case of rape reports: reclassifying rapes as "unfounded."

This strategy takes advantage of a socially constructed idea unique to rape: the common but erroneous belief that rape is prone to false allegations and unfounded complaints (Spohn and Horney, 2013). This allows police to alter rape statistics – unlike numbers for assaults, homicides, motor vehicle theft, and other serious crimes – by designating rapes "unfounded", a category reported in FBI data. In a comprehensive survey of problems in rape statistics, Yung (2014) finds many small studies in which police classified "ordinary rape complaints" with intoxicated victims as "unfounded", and evidence that police in large city departments used unfoundedness to reduce crime measures. "Indeed, police in Baltimore turned the UCR exception into a verb by openly stating that they had 'unfound' a rape complaint." (Yung, 2014)

FBI data on unfoundedness offer a national measure of data manipulation in rape statistics, but have not previously been used for this purpose to my knowledge. Since the analysis compares rates of "unfounded" classification before and after Compstat adoption, it is not necessary to assume that *all* "unfounded" rapes involve data manipulation. Rather, this strategy assumes that the rate of true (not manipulated) unfoundedness in reporting is continuous across the change in policy, and thus that Compstat-linked changes in the rate of unfoundedness indicate data manipulation.

To confirm that this effect measures data manipulation, not better record-keeping, I use a placebo test. I assess the effect of Compstat on the "unfounded" classification for other crimes, where the social justification for reclassification is absent. While true rates of false reporting are

Table 2: Unfoundedness and Manipulation in Crime Data

	Founded	*Unfounded*
Manipulated	Crimes manipulated in ways other than unfounding: downgrading to a less serious charge, suppressing victim reports, reclassifying to a misdemeanor, reducing the value of stolen property	Crimes manipulated by categorizing them as unfounded: social justification available when false reports are believed to be common
Not Manipulated	Crime recorded in crime statistics as it was reported to the police	Crime reported that truly did not occur: auto theft reports when the owner dumped the car, false theft reports for insurance fraud purposes, false reports of assault or rape, claims that a natural death was a homicide.

difficult to estimate, false reporting certainly does occur for other crimes. For motor vehicle theft, reports for the purpose of insurance fraud are common enough that the National Insurance Crime Bureau offers a training for law enforcement on identifying fraudulent reports (*Vehicle Theft Fraud*, N.d.). However, public and academic discourse around false reporting focuses overwhelmingly on sexual violence, giving police little justification for discretionary changes to other crime reports. Compstat-linked changes to "unfoundedness" of other crimes would raise concerns that increasing unfoundedness indicates better record-keeping, not data manipulation.

For both homicide and auto theft, some reports are declared unfounded. Unfounded reports need not inherently indicate manipulation: rather, reports may be founded or unfounded, manipulated or unmanipulated. Crimes may be manipulated without being labeled unfounded, as for example when crime reports are downgraded, suppressed, or reclassified. They may also be labeled unfounded without being manipulated, as in false theft or arson reports for insurance fraud (see Table 2). Rape is unique in that unfoundedness is an available tool for manipulation, and thus changes in unfoundedness around Compstat adoption give information about the level of manipulation.

4 Data and Analysis

4.1 New Data on Compstat Adoption

I analyze an original data set collected from newspapers, police department websites, and other public sources on the adoption of Compstat or similar metrics management techniques by police departments in 55 of the 68 largest cities in the United States. Details of the coding, sample, and measurement issues may be found in Appendix C (p. 5).

I add data on arrests and offenses known to police from the FBI's Uniform Crime Reporting Program, 1990-2013. I operationalize "major crimes" as crimes reported in Part 1 of the Uniform Crime Reporting guidelines (sometimes called index crimes), including homicide, manslaughter, rape, robbery, aggravated assault, and others (for details, including information on missingness, see Appendix C, p. 5). Many analyses of crime focus on arrests as a proxy for the number of crimes; however, these may vary separately, and their differences provide an important source of information. I operationalize minor crimes as crimes reported under Part 2 of the Uniform Crime Reporting guidelines, including arrests for drug charges, simple assault, fraud, stolen property, vandalism, weapons charges, prostitution, and quality of life offenses like drunkenness and disorderly conduct. Eight departments were dropped due to missing UCR records over a large portion of this period, for a total of 47 departments over 23 years, or 1081 unique department-year observations. Appendix A (p. 1) lists the cities included in the analysis, and which major variables are missing from each city. I also use FBI data on the number of offenses known to police and the number of clearances (how many crimes result in an arrest or other clear designation of the perpetrator).

4.2 Analytical Strategy

I use fixed-effects regression at the city-year unit of analysis to test the effect of metrics management on multiple outcomes measuring increases in minor arrests and data manipulation, as measures of *guaranteed* work. Including city-level fixed effects means that coefficient estimates capture variation *within* rather than *between* cities, so these results are not biased by unobserved differences between cities.

The estimated equation is shown below:

$$Y_{it} = \gamma Compstat_{it} + \beta X_{it} + City_i + Year_t + \epsilon$$

$Compstat_{it}$ is a binary independent variable describing whether a city has adopted Compstat; it is equal to zero until an agency adopts Compstat, and equal to 1 thereafter. X_{it} is a vector of control variables described for each individual analysis. $City_i$ is the agency effect, and $Year_t$ represents the year effect.

The main threat to the validity of the estimate is omitted variable bias. With city-level fixed effects included, coefficient estimates control for both observed and unobserved differences between cities. I also include year fixed effects, because both crime and arrests have substantial temporal variation. All standard errors are calculated using the wild bootstrap with the Rademacher distribution, which has good properties when the number of clusters is too small to achieve good performance with cluster-robust standard errors (MacKinnon and Webb, 2018).

I test the regressions with and without demographic covariates (percent of population that is Black/white) at the county level to control for variation across time within cities. The timescale runs across multiple data sources with different measurement techniques, and the estimation technique requires single-year estimates for appropriate controls, limiting the availability of demographic data (see Appendix C, p. 7). Using both geographic and temporal controls means coefficients are within-city differences, limiting concern about bias.

Still, it is impossible to fully exclude the possibility of an unmeasured time-varying feature of cities, perhaps a political change that coincides with the adoption of Compstat. Compstat might be rolled out during periods of conflict over high crime, when there is demand for harsher enforcement, or during transitions from more liberal to more conservative city politics. Appendix D (p. 8) includes many robustness checks, including an analysis addressing concerns about mayoral partisanship. There is no evidence of systematic trends in crime in the years preceding Compstat adoption.

Compstat in New York – the most commonly analyzed case – was controversial, and adopted during a period of exceptionally high crime locally and nationally. In journalistic and municipal records, there is little evidence that such politicization was common. Typically, articles describing Compstat adoption are short, factual, and contain mild commentary from residents and police. When Austin, TX, adopted Compstat, the "president of the Austin Police Association said he is waiting to see how Compstat will work and if it will reduce crime" (Plohetski, 2008). An article on Tulsa is typical, describing the new program briefly, as a new city technology. Studies of Compstat adoption emphasize that it became the standard technology for police management nationwide. Adoption was largely a matter of modernization and technical diffusion, not strong localized politics (Willis, Mastrofski and Weisburd, 2007). Despite the controversy associated with Compstat in New York, most cities do not have major political or crime-related changes driving the use of metrics management.

4.3 Results: Minor Arrests

I measure minor arrests using offenses reported under Part 2 of the UCR, like drug charges, vandalism, and quality of life offenses.[3] Because this data covers arrests, it is not subject to the limitations of data on "crime", which covers only crimes reported to and documented by police. The simplest way to increase recorded arrests is simply to arrest more people (Moskos, 2009).

In cities that adopt Compstat, arrests for minor offenses increase, both in absolute terms and as a share of the total number of arrests. Table 3 reports the effect of Compstat on the number of Part 2 arrests, while Table 4 reports the effect of Compstat on the share of arrests for Part 2 offenses. A jackknife analysis (in full: Appendix F, p. 17) shows that absolute results differ substantially with and without data from New York City. Including New York, adopting Compstat is associated with over 20,000 additional Part 2 arrests in a given city-year, or an increase of 1.3

[3] These crimes create disorder, which many would prefer to see addressed. Using this measure of "minor" crimes does not imply they are unimportant. Rather, these are highly discretionary arrests, where the "correct" threshold for arrest is least clear.

percentage points in the share of Part 2 arrests; excluding New York, this drops to over 3,500 arrests (without losing statistical significance). The estimate for the share of Part 2 arrests does not show this dramatic difference. Including NYPD in the analysis identifies the overall historical consequences of Compstat adoption, while the increase of 3,500 arrests is a better estimate for the consequences of adopting Compstat in cities other than New York. The share of arrests for Part 2 offenses increased by 1.3 percentage points, with or without New York.[4] Both analyses confirm that metrics management is not merely a tool enforcing accountability and preventing agents from shirking; it has substantive effects on the character of the policies being implemented.

Table 3: Effect of Compstat on Number of Part 2 Arrests

	Part 2 Arrests			Part 2 Arrests (excluding NYC)		
	(1)	(2)	(3)	(4)	(5)	(6)
Compstat	20,471.060**	20,294.410**	20,176.530**	4,028.177*	3,559.404*	3,581.072*
	(6,303.675)	(5,752.315)	(5,354.922)	(1,805.347)	(1,796.407)	(1,706.314)
Part 1 incidents	−0.304**	−0.209**	−0.206**	0.042**	0.036**	0.039**
	(0.019)	(0.019)	(0.019)	(0.007)	(0.007)	(0.007)
Total population		0.823**	0.796**		0.166**	0.125**
		(0.052)	(0.053)		(0.028)	(0.028)
Black population (%)			398,618.400*			320,364.200**
			(190,000.500)			(59,901.810)
White population (%)			94,309.340			19,492.110
			(172,207.300)			(54,041.780)
Observations	1,081	1,081	1,081	1,058	1,058	1,058

Note: †p<0.1; *p<0.05; **p<0.01
All regressions include year and agency fixed effects.

These results are substantively and statistically significant, accounting for 12% of the median Part 2 arrests in a city-year (or more if New York is included). These arrests may be for minor offenses, but they have serious consequences for the lives and communities of those arrested. Misdemeanor arrests lead to further entanglements in the criminal legal system, lost jobs, civic disengagement, and damaged community trust (Kohler-Hausmann, 2013; Goffman, 2014; Pinto,

[4] That is, NYPD skews the number of Part 2 arrests primarily because the city and its police force are so large, rather than because the patterns are markedly different.

Table 4: Effect of Compstat on Share of Part 2 Arrests

	Share of Part 2 arrests		
	(1)	(2)	(3)
Compstat	0.013^{**} (0.004)	0.013^{**} (0.004)	0.010^{*} (0.004)
Total population		0.000 (0.004)	-0.000 (0.004)
Black population (%)		0.755^{**} (0.144)	0.886^{**} (0.140)
White population (%)		$0.224^{\dagger}$ (0.126)	0.096 (0.118)
Part 1 incidents per capita			-0.650^{**} (0.078)
Observations	1,023	1,023	1,023

Note: †p<0.1; *p<0.05; **p<0.01

All regressions include year and agency fixed effects.

2015). They also, as I discuss below, interact with data manipulation to deny individuals and communities access to formal legal systems.

Because Compstat is associated with order maintenance policing, some might explain this as a result of adopting Compstat simultaneously with new policing goals. However, the median date of Compstat adoption in this data set is 2003, over 20 years after Kelling and Wilson's 1982 article on Broken Windows. Order maintenance policing is widely documented across a long period before most cities adopted metrics management. Moreover, journalism and case studies show that police adopt Compstat with the twin public goals of reducing violent crime and allocating resources more efficiently (Willis, Mastrofski and Weisburd, 2007; Marshall, 2009). Despite the association between Broken Windows policing and Compstat in New York, these philosophies spread across the country separately.

In terms of the model, the claim that Compstat was intended to increase arrests is equivalent to a preference among Supervisors for $guaranteed$ over $longterm$ work. The key insight of the model, though, is that even when Supervisors prefer that Agents focus on complex uncertain projects, they cannot reduce shirking without increasing the payoff of $guaranteed$ relative to $longterm.$

4.4 Results: Data Manipulation

I use "unfounded" classifications for reported rapes to test for data manipulation. Adopting Compstat is associated with an increase of 1.9 percentage points in the share of rapes designated unfounded (see Table 5). Since the mean share of rapes reported unfounded is 6.8%, adopting Compstat is associated with a 28% increase in the share of rapes reported as unfounded. Some may worry that Compstat would affect the number of rapes committed or reported to police, leading to post-treatment bias since the share of rapes declared unfounded is conditional on the denominator reported. Appendix D (p. 7) confirms that this analysis is substantively identical when the outcome variable is the total number of unfounded rapes.

Alternatively, Compstat might improve record-keeping: officers held accountable for their activities might need better records of case dispositions. To allay this concern, I use unfoundedness rates in auto thefts and homicides as placebo tests. If “unfoundedness” classifications increase because of better record-keeping, this effect should be visible across multiple types of crimes. These placebos resolve different problems.

Homicides are among the most difficult statistics to manipulate, since concealing them requires concealing murder victims. Thus, the “unfounded” classification of homicides should be hard for officers to manipulate. However, about 5.0% of homicides across years are classified as unfounded: if Compstat improves record-keeping, the share of homicides classified as “unfounded” should change.

For auto theft, in contrast, there are incentives for false reports, which are used in insurance fraud. Despite this incentive, the socially constructed justification available for “unfounding” rape cases is absent in auto theft. Police will not expect to be able to reclassify auto theft reports as “unfounded” to reduce the apparent crime rate. Unfoundedness changes for auto theft would suggest that record-keeping, not data manipulation, explained the change in rape classifications.

These placebo tests confirm that unfoundedness is not changing as a result of a general change in record-keeping, using two crimes with diverging incentives and social contexts. Meanwhile, changes in unfoundedness for rape provide a rare opportunity to assess data manipulation on a national scale.

There is no meaningful effect of adopting Compstat on the share of auto thefts or murders declared unfounded. Standard errors are large in comparison to the estimates, and point estimates are an order of magnitude smaller than those for rape. Results using counts of unfoundedness as the outcome variable are substantively identical. This suggests that data manipulation, not improved records, explains the change in unfoundedness for rape.

Clearances are another important outcome for police. Appendix B (p. 1) analyzes offenses known to police and clearances, plus data on arrests and Compstat adoption. The share of serious offenses and homicides cleared increases when departments adopt Compstat, but increased unfoundedness accounts for over half the improvement in rape clearance rates. Rape unfoundedness is unique in that manipulation leaves traces in the reported data, not in the incentives to

Table 5: Effect of Compstat on Share of Rapes Declared Unfounded

	Share Unfounded					
	Rape		Auto Theft		Murder	
	(1)	(2)	(3)	(4)	(5)	(6)
Compstat	0.020^{**} (0.005)	0.016^{**} (0.005)	0.005 (0.003)	0.003 (0.003)	−0.006 (0.005)	−0.007 (0.005)
Total population		-0.000^{**} (0.000)		-0.000^{*} (0.000)		−0.000 (0.000)
Black Population (%)		$0.311^{\dagger}$ (0.176)		$0.201^{\dagger}$ (0.107)		0.196 (0.157)
White Population (%)		−0.242 (0.154)		0.366^{**} (0.091)		$0.246^{\dagger}$ (0.138)
Observations	900	900	900	900	900	900

Note: †p<0.1; *p<0.05; **p<0.01
All regressions include year and agency fixed effects.
8 cities report no data on unfoundedness, and are excluded from this analysis.

alter reported data. Data manipulation for other offenses takes different forms, less susceptible to large-scale evaluation. Police may reclassify felonies as less serious crimes (Rashbaum, 2010; Poston and Rubin, 2014), as in the NYPD practice of recording sexual violence against sex workers as "theft of services" (Vogt, 2018); classify death investigations as "non-criminal" rather than homicides (Bernstein and Isackson, 2014); or discourage reporting and recording (Rayman, 2013). Police can also manipulate clearance and performance data by making arrests without adequate justification: generally, police face no repercussions for arresting the wrong person, and the prevalence of plea bargaining and prosecutor discretion means that innocent people sometimes plead guilty to avoid the risk of going to trial (Pfaff, 2017).

These forms of manipulation are largely invisible in UCR data. Often, they can only be identified by reinvestigating crimes and reinterviewing witnesses (Rashbaum, 2010; Poston and Rubin, 2014; Rayman, 2018). This labor-intensive process is infeasible for a national analysis. It also raises concerns about the validity of the increase in clearance rates: both increased arrests and improved clearance rates could result from data manipulation by police. The findings in this section can provide only a lower bound – probably much lower than the true value – for the level of data manipulation involved. Scholars and the public should treat administrative data with caution, as measures of internal processes as much as of the facts described.

4.5 Results: Crime Rates

Advocates for intensive policing argue that police focus on disorder reduces serious crime (Wilson and Kelling, 1982); more generally, advocates for metrics management hope that evaluating bureaucratic inputs will lead to changes in outputs. Research on the reasons departments adopt Compstat suggest that departments adopt Compstat to reduce serious crime, and improve departmental control over field operations, rather than to increase minor arrests (Weisburd et al., 2004). Indeed, advocates for Compstat focus largely on serious crime reduction (Henry, 2002). Similarly, proponents of other forms of metrics management argue that key outcomes for which bureaucracies are responsible – like educational achievement and public safety – will be most effectively reached with a bureaucracy focused on quantitative targets. I therefore test a final hypothesis: adopting Compstat will reduce serious crime.

Table 6: Effect of Compstat on Number of Serious Incidents and Part 1 Arrests

	Part 1 Incidents		Part 1 Arrests			Part 1 Arrests (excluding NYC)		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Compstat	214.511	-113.890	2,422.845†	2,334.490†	2,359.768*	391.243	91.104	93.064
	(3,282.051)	(3,193.718)	(1,247.250)	(1,230.400)	(1,006.193)	(705.274)	(670.120)	(669.008)
Demographics	No	Yes	No	Yes	Yes	No	Yes	Yes
Population	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Number of Part 1 Incidents	No	No	No	No	Yes	No	No	Yes
Observations	1,081	1,081	1,081	1,081	1,081	1,058	1,058	1,058

Note: †p<0.1; *p<0.05; **p<0.01
All regressions include year and agency fixed effects.

Arrests for serious offenses increase, consistent with the finding that clearances for serious crime increase. If reformers are correct that Compstat reduces serious crime, though, cities that adopt Compstat should see a reduction in offenses known to police. I find no such reduction, regardless of whether I examine index crimes together or consider murders, manslaughters, rapes, or other specific serious crimes. Coefficient estimates are close to zero, in dramatic contrast to the estimates for changes in Part 2 crimes, and well within the ranges for testing equivalence suggested by Hartman and Hidalgo (2019). This tracks with findings from O'Brien and Sampson (2015) that violence originates in private conflict rather than public disorder.

Departments with Compstat see more arrests for serious crimes and more murder arrests, which I discuss in more detail in Appendix B (p. 2). However, these results are driven entirely by New York City: without New York, the coefficient is entirely statistically insignificant, and the magnitude drops to as few as 91 Part 1 arrests. With or without New York, the increase in Part 2 arrests is about 3 times the increase in Part 1 arrests. Including New York, Part 2 arrests increase by 43% while Part 1 arrests increase by 17%; excluding New York, Part 2 arrests increase by 10% while Part 1 arrests increase by 3%.

Did this New York-specific increase improve service provision? Part 1 arrests may be valuable police services to combat serious crime. Or, they may result from a form of data manipulation: arrests without evidence or, sometimes, a clear underlying incident. The NYPD has been accused in multiple cases of pursuing arrests and charges in order to clear cases without sufficient evidence, casting doubt on the value of this rise in Part 1 arrests (McKinley Jr. and Baker, 2017). My analysis does not distinguish between these two interpretations. I urge caution in interpreting changes in clearance rates, both with respect to Compstat and more broadly.

Data manipulation raises more general concerns about using reported crime rates to measure the effects of changes in policing. Crime rates are a function not only of crimes committed, but also of civilian reporting, police recording, and publication practices. Changes in arrests are more reliable, since they are direct measures of police activity rather than proxy measures of external information.

5 Significance

5.1 Rethinking Bureaucratic Accountability

Metrics management influences policy implementation by street-level bureaucrats, substantially increasing minor arrests and data manipulation, and possibly increasing clearances. However, metrics management does not reduce the incidence of major crimes, as advocates hoped. The same incentive problems apply to other bureaucracies, suggesting that widespread use of metrics management creates incentives for data manipulation and a focus on $guaranteed$ work more broadly. The crucial problem for bureaucratic supervisors is to distinguish between shirking and stalled work. They can then reward officers who pursue challenging, valuable projects – even when those projects do not translate into crime statistics or arrests – and reduce incentives for data manipulation and minor arrests.

The formal model presented in this paper suggests one way to improve bureaucratic investment in complex, longterm work with probabilistic payoffs. Agents' preferences are set by the interaction between monitoring and their own valuations of the outcomes. Thus, bureaucratic supervisors could focus on hiring agents who are deeply motivated to pursue meaningful outcomes, or on creating bureaucratic cultures that celebrate and promote complex, meaningful accomplishments (Moynihan and Pandey, 2007). Drawing on Miller (1992) and Kreps (1990), supervisors might focus on cultures of trust and honor, rather than on supervision and monitoring.

Bureaucratic supervisors can also supplement quantitative data in two ways. First, supervisors at all levels might examine the data production process in detail to guard against data manipulation. Second, supervisors can look for outcomes that are challenging to manipulate, or contextualize quantitative measures with expanded qualitative data. However, any measure that becomes a core element of performance assessment is vulnerable to manipulation – and unfortunately, elected officials and other supervisors are not always averse. Instead, journalists often find evidence that elected officials seek to alter crime data to improve their prospects (Bernstein and Isackson, 2014). Researchers and journalists play a crucial watchdog role in assessing statistics on crime and bureaucratic performance.

Metrics management can be a valuable tool: it can prevent undesirable behaviors, when those are monitored (Mummolo, 2015), and reduce shirking. Metrics management is less effective at increasing work that is hard to monitor, especially when the details and quality of the work are essential to its success. Many key aspects of policing – solving crimes, defusing conflicts, and building civic trust – fall into this latter category, as do other important types of bureaucratic work. Both bureaucrats and researchers must attend to what particular types of data do and do not show.

In certain cases, metrics management actually reduces citizens' ability to monitor and evaluate state activity by leading to data manipulation, undermining the validity of the measures themselves, with "consequences for the administration of justice that may interfere with the legality and stated aims of law enforcement" (Skolnick, 1966). As these results show, metrics managers have not treated this loss of validity with sufficient seriousness.

Data manipulation is a critical methodological issue for scholars analyzing administrative data. Crime statistics are jointly produced by the criminal behavior of individuals, the enforcement activity of police officers, the reporting behavior of civilian, and the reporting behavior of officers. Researchers using bureaucratic data should treat it as a product of state activity rather than as a proxy for underlying behavior, and should be attentive to the problems created by policy changes around data collection. Metrics management itself alters the relationship between the distribution of criminal behavior and crime statistics by rewarding officers for particular data outcomes.

5.2 Metrics Management and Policy Feedback

In the case of policing, metrics management changes not only data and accountability, but the relationship of citizens to the political and legal system. Adopting Compstat is associated with at least 3,500 more arrests per city-year, a substantial rise in carceral contact compared to the median of 28,625 arrests. Misdemeanor arrests have important social and political consequences. They bring more people under state surveillance, put those arrested at risk for future warrants or harsher sentencing (Kohler-Hausmann, 2013), and impose economic losses (Pinto, 2015; Pager, 2007).

Contact with police carries physical risk even when the crime in question is minor (Eckhouse, 2019). Carceral contact leads to important political spillover effects: less voting, fewer requests for help from city government, less civic engagement (Burch, 2013; White, 2019). The burden of these minor arrests falls most heavily on Black and Latino communities, and especially on young men: the resulting distributive and procedural justice concerns damage the credibility of the criminal legal system and reduce cooperation with the law (Sampson and Bartusch, 1998; Fagan, 2008). The feedback effects of metrics management change who votes, who participates, who seeks and receives legal enforcement, and ultimately who has access to both justice and legal power.

Together, the increase in minor arrests and the downgrading of serious crimes have important effects for equal access to the law in the context of selective increases in enforcement of minor crimes. Minor arrests alienate those targeted from law enforcement and civic life. Fearing the state makes people with criminal records potential targets of violence: Goffman's informants in Philadelphia faced robbery precisely because their legal entanglements made them unwilling to seek police assistance. These same individuals, unable to seek police protection, had access only to retributive violence (Leovy, 2015; Goffman, 2014).

At the same time, data manipulation means crime victims lose access to the formal legal system: their victimization goes uncounted, and police do not effectively address it. In American cities, neighborhoods with high levels of violent crime are also neighborhoods with elevated arrest rates for minor crimes, where many serious incidents go unsolved (Goffman, 2014; Leovy, 2015). The dynamics described above offer one potential explanation. Increased enforcement of minor crimes separates citizens from the state. When alienation combines with the difficulty of getting police services after a violent incident, people turn to extralegal sources of justice, increasing the distance between these citizens and the state.

Taking metrics management seriously requires rethinking bureaucratic accountability.

References

Abrams, Lisa M, Joseph J Pedulla and George F Madaus. 2003. "Views from the classroom." *Theory into practice* 42(1):18–29.

Allen, David N. 1982. "Police supervision on the street." *Journal of Criminal Justice* 10(2):91–109.

Andersen, Simon Calmar and Donald P Moynihan. 2016. "Bureaucratic investments in expertise: Evidence from a randomized controlled field trial." *The Journal of Politics* 78(4):1032–1044.

Anderson, Elijah. 1978. *A place on the corner*. Chicago: University of Chicago Press.

Andrisani, Paul J, Simon Hakim and Emanuel S Savas. 2002. *The new public management*. Springer Science & Business Media.

Aviv, Rachel. 2014. "Wrong Answer." *The New Yorker* .

Bernstein, David and Noah Isackson. 2014. "The Truth About Chicago's Crime Rates." *Chicago Magazine* .

Bianco, William T, Peter C Ordeshook and George Tsebelis. 1990. "Crime and punishment: Are one-shot, two-person games enough?" *American Political Science Review* 84(2):569–586.

Blau, Peter Michael. 1963. *The dynamics of bureaucracy*. Chicago.

Bovens, Mark, Robert E Goodin and Thomas Schillemans. 2014. *The Oxford handbook public accountability*. Oxford University Press.

Brayne, Sarah. 2014. "Surveillance and System Avoidance." *American Sociological Review* .

Brehm, John and Scott Gates. 1993. "Donut shops and speed traps." *American Journal of Political Science* pp. 555–581.

Brehm, John and Scott Gates. 1997. "Working, shirking, and sabotage." *Ann Arbor: University of Michigan Press* .

Burch, Traci. 2013. *Trading democracy for justice*. University of Chicago Press.

Carrigan, Christopher. 2018. "Clarity or collaboration: Balancing competing aims in bureaucratic design." *Journal of Theoretical Politics* 30(1):6–44.

Eberhardt, Jennifer L, Phillip Atiba Goff, Valerie J Purdie and Paul G Davies. 2004. "Seeing black." *Journal of Personality and Social Psychology* 87(6):876.

Eckhouse, Laurel. 2019. "Everyday Risk." *Working paper* . Accessed November 12, 2020.
URL: *https://docs.wixstatic.com/ugd/b323fb_8680236ca1aa4d7ea4193b6bb91ed03d.pdf*

Engel, Robin Shepard. 2000. "The effects of supervisory styles on patrol officer behavior." *Police Quarterly* 3(3):262–293.

Eterno, John and Eli B. Silverman. 2012. *The crime numbers game*. Boca Raton, FL: CRC Press.

Fagan, Jeffrey. 2008. "Legitimacy and criminal justice-introduction." *Ohio St. J. Crim. L.* 6:123.

Federal Bureau of Investigation Staff. 2012. Crime in the United States 2012: Uniform Crime Reports. Technical report. Accessed November 12, 2020.
URL: *https://ucr.fbi.gov/crime-in-the-u.s/2012/crime-in-the-u.s.-2012*

Gailmard, Sean and John W Patty. 2007. "Slackers and zealots." *American Journal of Political Science* 51(4):873–889.

Gailmard, Sean and John W. Patty. 2012. "Formal Models of Bureaucracy." *Annual Review of Political Science* 15:353–377.

Gelman, A., J. Fagan and A. Kiss. 2007. "An analysis of the New York City Police Department's stop-and-frisk policy in the context of claims of racial bias." *Journal of the American Statistical Association* 102(479):813–823.

Glaser, Jack. 2015. *Suspect race*. Oxford University Press, USA.

Goffman, Alice. 2014. *On the Run*. University of Chicago Press.

Goldstein, Joseph. 2014. "Forbidden Zone for the Police: Places Ready-Made for a Nap." *The New York Times* .

Goldstein, Rebecca, Michael Sances and Hye Young You. 2018. "Exploitative Revenues, Law Enforcement, and the Quality of Government Service." *Urban Affairs Review* .

Gruening, Gernod. 2001. "Origin and theoretical basis of New Public Management." *International Public Management Journal* 4(1):1–25.

Hartman, Erin and F. Daniel Hidalgo. 2019. "An equivalence approach to balance and placebo tests." *American Journal of Political Science* .

Henry, Vincent E. 2002. *The COMPSTAT paradigm*. Looseleaf Law Publications.

Holmstrom, Bengt. 1982. "Moral hazard in teams." *The Bell Journal of Economics* pp. 324–340.

Jeffers, Greg. 2014. Proactive Policing in a Majority Black, Urban Community PhD thesis IU, Bloomington: . Accessed November 12, 2020.
URL: *https://search-proquest-com.du.idm.oclc.org/docview/1564775725?accountid=14608*

Kelley, Judith G and Beth A Simmons. 2015. "Politics by number." *American journal of political science* 59(1):55–70.

Kohler-Hausmann, Issa. 2013. "Misdemeanor Justice." *American Journal of Sociology* 119(2):351–393.

Kreps, David M. 1990. "Corporate culture and economic theory." *Perspectives on Positive Political Economy* p. 90.

Langbein, Laura. 2010. "Economics, public service motivation, and pay for performance: complements or substitutes?" *International Public Management Journal* 13(1):9–23.

Leovy, Jill. 2015. *Ghettoside*. Spiegel & Grau.

Lipsky, Michael. 2010. *Street-level bureaucracy*. Russell Sage Foundation.

MacKinnon, James G and Matthew D Webb. 2018. "Randomization inference for difference-in-differences with few treated clusters." *Queen's Economics Department Working Papers* .
URL: *http://qed.econ.queensu.ca/working_papers/papers/qed_wp_1355.pdf*

Marshall, Nicole. 2009. "Tulsa police unveil new CompStat crime center." *The Tulsa World* .

McCubbins, Mathew D., Roger G. Noll and Barry R. Weingast. 1987. "Administrative Procedures as Instruments of Political Control." *Journal of Law, Economics, & Organization* 3(2):243–277.

McCubbins, Mathew D. and Thomas Schwartz. 1984. "Congressional Oversight Overlooked: Police Patrols versus Fire Alarms." *American Journal of Political Science* 28(1):165–179.

McKinley Jr., James C. and Al Baker. 2017. "Teenager Who Says Police Coerced Witnesses Faces Trial." *The New York Times* .

McLaughlin, Kate, Stephen P Osborne and Ewan Ferlie. 2002. *New public management: Current trends and future prospects*. Psychology Press.

Meier, Kenneth J and George A Krause. 2003. "The scientific study of bureaucracy: An overview." *Politics, policy, and organizations* pp. 1–19.

Meier, Kenneth J, JL Polinard and Robert D Wrinkle. 2000. "Bureaucracy and organizational performance." *American Journal of Political Science* pp. 590–602.

Miller, Gary J. 1992. *Managerial dilemmas*. Cambridge University Press.

Miller, Gary J. 2005. "The political evolution of principal-agent models." *Annu. Rev. Polit. Sci.* 8:203–225.

Moskos, Peter. 2009. *Cop in the hood*. Princeton University Press.

Moynihan, Donald P and Sanjay K Pandey. 2007. "The role of organizations in fostering public service motivation." *Public administration review* 67(1):40–53.

Mummolo, Jonathan. 2015. "Can New Procedures Improve the Quality of Policing.".

Murakawa, Naomi. 2014. *The first civil right*. Oxford University Press, USA.

O'Brien, Daniel Tumminelli and Robert J. Sampson. 2015. "Public and Private Spheres of Neighborhood Disorder." *Journal of Research in Crime and Delinquency* 52(4):486–510.

Pager, Devah. 2007. *Marked*. Chicago: University of Chicago Press.

Pfaff, John. 2017. *Locked in*. Basic Books.

Pinto, Nick. 2015. “The Bail Trap.” *The New York Times* .

Plohetski, Tony. 2008. “New police program emphasizes statistics, strategy, accountability.” *The Austin American-Statesman* .

Poston, Ben and Joel Rubin. 2014. “LAPD Misclassified Nearly 1200 Violent Crimes as Minor Offenses.” *Los Angeles Times* .

Rashbaum, William K. 2010. “Retired Officers Raise Questions on Crime Data.” *New York Times* .

Rayman, Graham. 2018. “NYPD probes claims cops fudging crime stats.” *New York Daily News* .

Rayman, Graham A. 2013. *The NYPD Tapes*. Macmillan.

Rios, Victor M. 2011. *Punished*. NYU Press.

Sampson, Robert J. and Dawn Jeglum Bartusch. 1998. “Legal cynicism and (subcultural?) tolerance of deviance.” *Law and Society Review* pp. 777–804.

Skolnick, Jerome H. 1966. *Justice Without Trial*. New York: Wiley.

Spohn, Cassia and Julie Horney. 2013. *Rape law reform*. Springer Science & Business Media.

Stuntz, William J. 2011. *The collapse of the American criminal justice system*. Cambridge, Mass.: Belknap Press of Harvard University Press.

Thompson, Cheryl W. 2000. “Police Often Close Cases Without Arrest.” *Washington Post* .

Tsebelis, George. 1989. “The abuse of probability in political analysis.” *American Political Science Review* 83(1):77–91.

Turner, Ian R. 2017. “Working smart and hard? Agency effort, judicial review, and policy precision.” *Journal of Theoretical Politics* 29(1):69–96.

Vehicle Theft Fraud. N.d. Accessed November 12, 2020.
URL: *http://www.nicbtraining.org/vt/index.html*

Vogell, Heather. 2011. “Investigation into APS cheating finds unethical behavior across every level.” *Atlanta Journal-Constitution* .

Vogt, PJ. 2018. “The Crime Machine, Parts I and II.” *Reply All* (128-129).

Walker, Hannah. 2018. “Targeted.” *The Journal of Politics* .

Weaver, Vesla M. 2012. Embedding Crime Policy. In *Living Legislation*, ed. Jeffery A. Jenkins and Eric M. Patashnik. University of Chicago Press.

Weisburd, David, Stephen D. Mastrofski, Rosann Greenspan and James J. Willis. 2004. *The growth of Compstat in American policing*. Police Foundation Washington, DC.

White, Ariel. 2019. “Misdemeanor Disenfranchisement.” *APSR* .

White, Michael D. 2008. “Identifying Good Cops Early Predicting Recruit Performance in the Academy.” *Police Quarterly* 11(1):27–49.

Willis, James J., Stephen D. Mastrofski and David Weisburd. 2007. “Making sense of COMPSTAT.” *Law & Society Review* 41(1):147–188.

Wilson, James Q. 1978. *Varieties of Police Behavior*. Harvard University Press.

Wilson, James Q. 1989. *Bureaucracy*. Basic Books New York.

Wilson, James Q and George L Kelling. 1982. “Broken windows.” *Atlantic Monthly* 249(3):29–38.

Yung, Corey Rayburn. 2014. “How to Lie with Rape Statistics.” *Iowa Law Review* 99(1197).

Zacks, Richard. 2012. *Island of Vice*. Anchor.