A FEW BAD APPLES: AN EMPIRICAL ASSESSMENT OF RACIAL DISPARITIES IN POLICING

By

Travis M Carter

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Criminal Justice – Doctor of Philosophy

2024

ABSTRACT

Although ample research has focused on the identification of racial disparities in policing, minimal research has focused on the remediation of those disparities once identified in police agencies. To instigate evidence-based police reform, scholars must begin to consider both the identification and remediation of racial disparities in policing. This dissertation provides a first step in that direction by analyzing detailed data from the Chicago Police Department (CPD) from 2012 to 2015 through the use of innovative statistical models and quantitative techniques to determine which police officers engaged in racially-disparate stop and arrest behavior after accounting for when, where, and under what circumstances they conducted their daily police work. Two sets of policy-relevant questions were then explored to better understand the viability of evaluating actual policy interventions that seek to remediate disparities in policing. The first question has its roots in diversification and seeks to understand whether officers from under-represented identities in policing engage in more racially equitable police behavior than officers with over-represented identities. The second question stems from situational crime prevention and explores whether racially disparate police behavior is influenced by the presence (or lack of) of situational opportunities to contact citizens with racial minority identities. Results showed that racial disparities in stops and arrests were concentrated among 5% of the entire sample of CPD officers. Officer demographics did not consistently predict whether an officer had engaged in racially disparate behavior. Moreover, there was little evidence to suggest that officers from under-represented racial and gender identities engaged in more (or less) racially equitable stops and arrests when compared to their White and male colleagues. However, officers' racially-disparate behavior was influenced by situational factors related to where they work and who they work with. The implications of these findings for future research and police reform are discussed in detail.

ACKNOWLEDGEMENTS

The most important lesson that I learned on my path toward earning a doctorate degree is that you cannot walk it alone. I have many to thank for this lesson.

First, I would like to thank my dissertation committee—Drs. Scott Wolfe, Jeff Rojek, Chris Melde, and Justin Nix. Their continued support is invaluable, and their wisdom is immeasurable. Countless times they opened their doors to me, which I would walk through both ways with more questions than answers. They helped me see research for what it should be: a scientific inquiry ripe for critical assessment and continued investigation.

While serving as my dissertation Chair, Dr. Wolfe helped me learn how to forge my path. Dr. Wolfe was also my mentor – and I will cherish our work for the rest of my life. Working with Dr. Wolfe also helped me realize that our research ethos shapes the path we as scholars walk for the rest of our careers. That path, no matter how winding it may be, will always focus on doing important research *with* and *for* criminal justice practitioners. I thank him for helping me find that ethos.

Second, I would like to recognize the faculty members in the School of Criminal Justice (SCJ) who helped me lay the groundwork for my path and charged my research ethos. Drs. Steve Chermak, Jeff Rojek, and Edmund McGarrell were the first to introduce me to "action research" and criminal justice theory. Their teachings will live on in my research for years to come. Drs. Tom Holt and Chris Melde, whether they realized it or not, helped me develop new theoretical interests and learn how to navigate the reality of research, being one with many paths rather than just one.

Third, I would like to thank my SCJ cohort – Amanda Osuna, Brenna Helm, John Ropp, Noah Turner, and Stephen Oliphant – we all started at square one and while our paths have

diverged in unique and exciting ways, I will always be thankful for our shared memories. To my past and present SCJ colleagues – Alyssa LaBerge, Ariel Roddy, David (Yongjae) Nam, Jedidiah Knode, Jin Lee, Jordan Parker, Makayla Burden, Spencer Lawson, Sydney Litterer, and Yang (Vincent) Liu – it has been a pleasure to walk this path alongside you all.

Lastly, I would like to thank my dearest family and friends. I would like to thank my mom – Christie Wells-Paddock – for teaching me to follow my heart down the path that it leads me and to enjoy that path as I walk along it. I would like to thank my dad – Michael Carter – for teaching me to stay true to my path no matter the challenges that lie ahead. To my best friends – Alex Burhop and Gianluca Guadagno – their support and loyalty are irreplaceable. Finally, I would like to thank my dear Sylvia for everything. She above all has kept me on my path. Frankly, I would not know where or how to walk without her. You are my world and I love you.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

Problem Statement

Policing in the United States is by nature a coercive process, involves immense discretion, and is open to public scrutiny. Although democratic accountability is vital to functional institutions, it can create ambiguous public expectations that generate goal conflict in police agencies. The implications of this are felt by police officers on the front line, making their decisions personally challenging and behavior publicly consequential. These consequences have been historically salient for racially marginalized members of society.

Some have called for abolishing the police in hopes of mitigating these consequences and, in doing so, eliminating the coercive nature of police work altogether (McDowell & Fernandez, 2018). Though with crime still a major concern, most Americans acknowledge a need for coercive authority and simply desire a higher quality of policing, instead (Lum, 2021; McCarthy, 2022). Some of the most popular reform proposals seek to elevate the quality of police services by improving police-citizen relations in local communities (McCarthy, 2022). A prime example is the early implementation of community policing strategies developed at the turn of the 1960s, such as the Flint Foot Patrol Project (Trojanowicz, 1983). Community policing was a paradigm shift both organizationally and socially for police agencies; designed to innovate police deployment strategies in a way that builds police-citizen relations from the ground up and creates downstream impacts on crime and fear of victimization (Reisig, 2010). Yet, even as well-intentioned as these strategies were, scholars have consistently articulated that the design, practice, and reward structure of policing in the U.S. has historically worked against these intentions for communities of color—so much so that they perpetually feel over-policed and over-punished (Cobbina-Dungy, 2019; Jones-Brown, 2007; Sierra-Arévalo, 2021).

It bears noting that the feeling of being over-policed is not an isolated experience among non-White Americans. Evidence consistently reveals racial disparities in police behavior across studies and settings (Neil & Winship, 2019). The salience of racially-disparate policing is magnified when considering its association with numerous public safety and public health-related outcomes for people of color. For example, police force and misconduct are more likely to occur in racially diverse communities (Hoekstra & Sloan, 2022; Terrill & Reisig, 2003). These disparate police behaviors are historically associated with greater distrust among marginalized community members (Kirk & Papachristos, 2011), which is linked to a reduced willingness to cooperate with the police to solve crimes (Brunson & Wade, 2019; Tyler & Fagan, 2008). Racially-disparate police contact is also associated with having worse mental and physical health among non-White adults (Bor et al., 2018; Geller et al., 2014; Sewell & Jefferson, 2016), and poorer educational performance among non-White youth (Ang, 2020; Legewie & Fagan, 2019). More generally, racial disparities in lethal police force simply lead to a higher rate of lives lost at the hands of the police among people of color (Wrigley-Field, 2020).

Research has dedicated considerable attention to identifying when racial inequalities in police behavior arise in police agencies and what mechanisms enable it (Smith & Alpert, 2007; Smith et al., 2006). Understanding what drives racially-disparate police behavior may help agencies improve the quality of their police services and prevent unintended consequences for people of color. This line of research continues to lead today's police reform efforts as calls to enhance police accountability (e.g., civilian review boards), training (e.g., cultural sensitivity), and more drastic reform measures (e.g., defunding) pervade political discourse (McCarthy, 2022).

The problem is, however, less empirical research focuses on how to address racially-disparate police behavior once identified in an agency. For example, diversification policies have

gained increased favor among the public in recent years given their correspondence with public interest in the enhancement of police-citizen relations (McCarthy, 2022). Interventions focusing specifically on racial diversification claim that enhancing an agency's racial representation with the public it serves can positively impact police-citizen interactions and calm growing public concerns over race relations (Theobald & Haider-Markel, 2009; Weitzer, 2015). Yet, recent empirical evidence providing support for racial diversification policies is limited (Ba et al., 2021; Goncalves & Mello, 2021; Hoekstra & Sloan, 2022).

Other policies, such as gender diversification, share similar goals and have received comparatively more empirical support over the years (Keiser et al., 2002; Meier & Nicholson-Crotty, 2006). Unfortunately, there is insufficient research evidence indicating whether these policy interventions effectively address racially-disparate police behavior. This is concerning given the increasing demand for evidence-based policing and calls for data-driven police reform (Ridgeway, 2018; Sherman, 2013), which indicates the need for more research that focuses on how to identify and respond to racially-disparate police behavior. Conducting more policy-focused research is thus critical to enhancing police legitimacy, police-citizen relations, and the public health and safety of communities that are most impacted by racially-disparate policing.

Current Study

The current study seeks to advance research on racial disparities in policing by analyzing data from police officer shift assignments, stops, and arrests in the Chicago Police Department between 2012 and 2015. In so doing, this study aims to make three contributions to the literature. The central contribution of this study is to provide an empirical assessment of racially-disparate police behavior engaged by individual officers using highly detailed information about their patrol assignments and activity. Most studies of racial disparity have been conducted at an aggregate

level with the assumption that such behavior is pervasive across the agency (Alpert, Smith, & Dunham, 2004; Gelman et al., 2007; Smith et al., 2021); while others have explored disparities between racial groups of officers in the aggregate (Ba et al., 2021). Yet, research has shown that police officer misconduct can be concentrated among a few officers in the agency (Chalfin & Kaplan, 2021; Christopher, 1991), and this can be true for racial disparities in traffic stop behavior as well (Ridgeway & MacDonald, 2009). This parallels research on the etiology of criminal offending, which has consistently shown that criminal behavior is concentrated among small subsets of the U.S. population (Gottfredson & Hirschi, 1990; Wolfgang, Figlio, & Selling, 1972).

Failing to account for the potential concentration of racially-disparate police behavior may lead to inaccurate conclusions about the extent and nature of its presence in an agency and generate ineffective policy interventions that seek to address them. Despite these concerns, research has been largely unsuccessful in capturing such information in empirical analyses (Neil & Winship, 2019). Accordingly, the first component of this study involves conducting an officer-level analysis using methods proposed by Ridgeway and MacDonald (2009) to create "customized internal benchmarks" for each officer in the agency. This will allow for the identification of which officers (if any) have engaged in a pattern of racially-disparate police behavior.

Until now, most research on racial disparities in policing focuses on identification, not remediation. As such, the second contribution of this study is to empirically test two sets of policy-relevant questions based on policies aimed at addressing racially-disparate behavior. The first policy-relevant question involves exploring whether officers with under-represented gender and racial identities engage in more racially equitable enforcement behavior than their male and White colleagues.

To answer this question, the following study draws on empirical evidence from previous

analyses to construct a dataset of at-risk officers for racially-disparate police behavior and compares the proportion of stops and arrests involving Black citizens made by each of these officers to those made by other officers in the agency who were not found to be at risk for engaging in racially disparate behavior. While these direct comparisons generate a disparity estimate for each at-risk officer, stops and arrests were matched similarly based on time, place, and context to ensure estimates were as statistically unbiased as possible. To assess whether demographic factors help explain differences in the racial equity of police behavior, stops and arrests made by not-at-risk officers were subset by officer race and gender, to see if variation in disparity estimates correspond systematically across racial and gender subsets.

The second policy-relevant question involves exploring whether the probability of an officer engaging in racially disparate police behavior is associated with the racial composition of the residential population within the police beat that they are working in on any given day. If officers who were previously identified as being at an increased risk of engaging in racially-disparate behavior have a lower chance of engaging in disparate behavior in police beats with smaller concentrations of Black residents, this might suggest that their disparate behavior is motivated in part by opportunities to contact them. Accordingly, evidence from these analyses will provide critical insight into the growing evidence base on mechanisms by which agencies may or may not solve their racial disparity problems.

Lastly, this study contributes to theory and practice by providing key insights into the challenges of estimating racial disparities in police behavior while also demonstrating novel ways to approach the situation when such data are available. In doing so, I provide new ways to estimate the effectiveness of policy interventions and improve upon the tools that scholars have used to estimate racial disparities so that they can be more transparent about the statistical certainty of

their findings.

Overview

The remainder of this dissertation is organized as follows. In Chapter 2, I begin by defining racial discrimination and racial disparity in policing and use these working definitions to organize a brief review of the etiology of each concept. I then document the evolving body of research that explores how best to identify racial disparities in police behavior, paying particular attention to the challenges to estimation and directions for future research. Next, I review the research on policy interventions aimed at reducing racial disparities and discrimination in policing. I document the barriers that police practitioners and scholars face when implementing and evaluating these interventions. The chapter concludes with a presentation of the conceptual framework that lays the foundation for the current study and its implications for the advancement of research on racial disparities in policing.

Chapter 3 follows by describing the data sources, data cleaning, and empirical strategies used in this study. In the first section, I outline each data source and how they were collected. The next section discloses what data were omitted from the finalized datasets and why. Justifications for the spatial level of analysis are provided as well. The final two sections outline the conceptual framework this study uses to identify racial disparity among individual officers and how it evaluates policy-relevant questions aimed at mitigating disparities when perpetuated by officers.

In Chapter 4, I describe the results from each step of the empirical strategy described above. The first section reports the results of the officer-level analyses of racially-disparate police behavior across stops and arrests. Before jumping into the policy questions, results will be reported on the dynamics of those who engaged in disparate behavior, whether there were any relevant predictors, and if the disparity for a given officer was consistent across stops and arrests. The next

section reports the results of some indirect evidence in favor of (or against) policy interventions based on a set of policy-relevant questions identified from the literature.

Chapter 5 concludes with a summary of the main findings and contributions of this study. Additionally, I describe the policy implications of these results while also acknowledging their potential limitations. I then conclude with a discussion on the paths moving forward for future researchers to follow, outlining how to build on this study, and how to extend racial disparity and discrimination research more generally.

CHAPTER 2: LITERATURE REVIEW

Defining the Problem: Racial Disparities and Discrimination

There is ongoing debate as to how researchers should define racial discrimination in the social sciences, which has important implications for the study of discrimination in policing. According to the National Research Council (2004, p. 39-40), racial discrimination involves two components:

> (1) Differential treatment on the basis of race that disadvantages a racial group and (2) treatment on the basis of inadequately justified factors other than race that disadvantages a racial group (differential effect).

Quantitative researchers are primarily interested in the first component of this definition because it allows them to empirically test the significance of civilians' race for predicting variation in police-citizen interaction-related outcomes. This is exemplified in Neil and Winship's (2019) annual review of tests of racial discrimination in policing, where they similarly define racial discrimination as occurring when "…similarly situated individuals of different races are treated differently" (p. 77).

Unlike the definitions above, I conceptualize racial discrimination in policing as involving an individual officer engaging in differential *behavior* towards a racial group and having a corresponding *intent* to do so (for similar conceptualizations, see Engel, 2008; Smith & Alpert, 2007; Tomaskovic-Devey, Mason, & Zingraff, 2004).[1] By accounting for individual officers' intent, scholars can begin to consider reasonably justifiable situations where racial differences in police behavior do not constitute discrimination. Indeed, the NAS (2018) describes the resulting racial differences in these situations as racial disparities, which are objective differences in police

[1] Intent is defined as an officer's willingness to treat a specific racial group differently than another racial group. The underlying nature of the intent, whether it is explicit or implicit, will be discussed at length in a later section.

behavior that correspond with civilian race but are not solely because of one's race.

For example, racial disparities in an individual police officer's behavior may be the result of corresponding racial differences in criminal involvement and geographic differences in where they are assigned to patrol (Tomaskovic-Devey et al., 2004). The fact that individual officers in these situations do not harbor an intent to treat people of color differently provides them with a reasonable justification for their actions and gives researchers important contextual information to better understand when this behavior could lead to discrimination. By shifting the focus to individual officers and accounting for context, I can more precisely determine when discrimination is exemplified through police behavior.

Failing to acknowledge the dynamic nature of individual officer intent when trying to understand racial differences in police behavior is not only a conceptual issue but also an empirical one. For example, differences in police behavior may be the result of differences in the racial composition of an offending population. If this is the case, officers' intention to behave may be morally sound, yet differences in their behavior remain. Acknowledging the conceptual distinction between discrimination and disparities will thus assist in the identification of their respective causes (Engel, 2008). While this sentiment underlies the motivations and criticisms on the validity of tests of racial discrimination in policing, still much of the literature defines racial discrimination as merely racially-disparate police behavior (Neil & Winship, 2019).

With a definition of racial disparity and racial discrimination established, the following section provides a brief overview of the proposed causes of racially-disparate policing. Thereafter, I discuss the proposed theoretical mechanisms that may generate a racially motivated *intent* behind such *behavior*, making it discrimination. Organizing the literature in this manner ensures that the causes of disparate police behavior are not confounded with what may uniquely contribute to racial

discrimination.

Proposed Causes of Racial Disparities in Policing

Macro-Level Explanations

Reviews on racial disparities in policing highlight six common causes, four of which focus on macro-level processes and two of which focus on micro-level processes (see Braga et al., 2019; NAS, 2018; Tomaskovic-Devey et al., 2004). Starting at a macro-level, one explanation is that the historical overrepresentation of non-White Americans in criminal offending created corresponding racial disparities in police encounters and arrests of non-White Americans (MacDonald et al., 2001; Sampson & Lauritsen, 1997). However, racial differences in criminal offending are concentrated primarily among the most serious offenses (see Braga et al., 2019).

For example, according to National Incident-Based Reporting System data from 2021, Black or African Americans comprise 12% of the U.S. population in 2021, yet they make up 44% of all known offenders involved in violent crimes and 12% of all known offenders involved in property crimes. Accordingly, while Black or African Americans are overrepresented among violent offenders relative to their racial composition in the U.S. population, they are equally represented among property offenders.

A second macro-level explanation is that racial disparities in policing are due to differential enforcement of crimes most likely to be committed by non-White Americans (Smith & Alpert, 2007; Tomaskovic-Devey et al., 2004). Police are more likely to enforce more serious offenses; therefore, Black or African Americans would be overrepresented in police contacts due to their differential participation in violent crime (NAS, 2018). This could be the case for other crime types as well.

One popular example involves the differential enforcement of crack versus cocaine in the

late 1980s and early 1990s. Drug arrest rates trended consistently across racial groups from the 1960s until the late 1980s, whereupon Black Americans were arrested at an increasingly disparate rate compared to White Americans (Blumstein, 1993). At about the same time, the Anti-Drug Abuse Act of 1988 created a 100:1 crack versus cocaine sentencing disparity that scholars have argued ultimately encouraged law enforcement to target street drug crimes, such as crack, in open-air markets, which were primarily facilitated by non-White Americans (Hagan, 2012). In contrast, cocaine—a drug more commonly used among White middle- and upper-class Americans—was less likely to be targeted by law enforcement given the lower consequences of possession and because it was more likely to be sold in closed markets (Blumstein & Beck, 1999; Tonry, 2011). Therefore, the differential enforcement of crack over cocaine could be a reason why racial disparities in drug arrests among Black Americans were increasing during that time. More generally, this suggests that racial disparities in police outcomes could be due to differential enforcement of crimes committed more so by Black Americans. Importantly, this represents only one example of differential enforcement as a result of differential criminal involvement—this may not always be the case. Accordingly, other explanations are needed to explain racial disparities in other criminal activities.

A third and related macro-level explanation is that many of the person-focused approaches to policing, such as "pulling levers" strategies outlined in Project Safe Neighborhoods and Ceasefire-inspired initiatives, differentially target high-activity offenders (NAS, 2018). Those most likely to be targeted by these initiatives are non-White Americans due, in part, to their higher rates of recidivism (Kubrin, Squires, & Stewart, 2007; Spohn & Holleran, 2002; Wehrman, 2010). As a result, disparate police behavior may be due to an overrepresentation among non-White Americans in the recidivist population, which is most likely to be targeted by proactive policing

strategies that utilize person-focused approaches to crime reduction.

Drawing from the communities and crime literature, a macro- and meso-level explanation suggests that community contexts may give rise to offending and racial disparities in police behavior (Tomaskovic-Devey et al., 2004). Given the concentration of violence in predominantly non-White and disadvantaged communities (Sampson et al., 1997), the amplified police presence in these communities can create more opportunities for police contact with non-White residents. The nature of these contacts is important as well. Residents from disadvantaged communities generally have worse perceptions of the police, which is associated with a lower willingness to work with them (Brunson & Wade, 2019; Tyler & Fagan, 2008). Police officers perceive people in these communities as more dangerous and suspicious (Klinger, 1997; Smith & Alpert, 2007), and perceive them to engage in more aggressive behavior (Rengifo & Fowler, 2016). Collectively, this generates more tension during interactions with the police and can create racial disparities in police use of force against people of color in disadvantaged and racially diverse communities (e.g., Hoekstra & Sloan, 2022; Terrill & Reisig, 2003).

Micro-Level Explanations

Shifting attention to a more micro-level, two complimentary explanations for racial disparities exist. In a sequence of two papers, the first by Smith and colleagues (2006) and the second by Smith and Alpert (2007), the authors reject the idea that racial differences in police behavior are due to race being at the forefront of officers' minds. Instead, they contend that there must be a non-volitional explanation for the disparate treatment of marginalized people (Smith & Alpert, 2007, p. 1269).

Accordingly, in their first theoretical explanation of racial disparities in police behavior, Smith et al. (2006) draw on Tomkin's (2008) script theory from cognitive psychology to develop

their theory of differential suspicion. The social construction of suspicion in policing can, however, more broadly be described by integrating insights from symbolic interactionism and cognitive psychology. As such, these theoretical frameworks operate under the shared assumption that officers are semi-autonomous agents with an immense latitude of discretion over whom to interact with; and have the capacity to be influenced (albeit not entirely determined) by external forces. Much of the literature until now has shown indirect support of this, either by suggesting that formal rules only go so far in controlling officer behavior (Bittner, 1967), or by contending that officers interpret rules, norms, and cultural expectations as they see fit in their working environment and mete out their own bureaucratic policy (Crank, 1998; Lipsky, 1980).

How police officers operate in their working environments is dictated in no small part by how they perceive information and learn how to interact in those environments. Outside the academy, a critical path toward defining and refining suspicion is through officers' interactions with people. Officers must rely on perceived information to make judgments about people's morality and potential criminality to preserve safety and 'maintain the edge' over them (Muir, 1977; Van Maanen, 1978). Police-citizen interactions serve as opportunity contexts, which reveal to officers crucial information about potential suspects in terms of the real or perceived consequences of peoples' actions (Blumer, 1986; Mead, 1934; Thomas & Thomas, 1928). This information-gathering process is contextualized within the environment in which these interactions exist and then translates into how officers make risk-informed decisions about suspicion within those interactions and environments (Klinger, 1997). Policing scholars have more broadly characterized this process of defining suspicion as how officers look for cues (i.e., symbols) of suspicious behavior, whether it be their symbolic assailants (Skolnick, 1966), their ungovernables (Muir, 1977), or their assholes (Van Mannen, 1974).

Based on this information, Smith and colleagues (2006) draw on script theory from cognitive psychology to better understand how these loose connections between symbols of suspicion can translate into actualized behavior by police officers (Tomkins, 2008). They suggest that the scripts officers develop over the course of their interactions with civilians provide them with a set of rules for how to interpret, predict, and control situations. These scripts are predicated on repeated events that store loose networks of information such as symbols of suspicion to interactional outcomes that become thicker and more systematic cognitive schemas over time (Smith & Alpert, 2007).

The cognitive schemas that police officers develop help form a mental image of who is suspicious, which can influence subsequent behaviors when officers identify familiar characteristics that correspond to those schemas. As a result, how officers characterize suspicion can lead to different police outcomes. Smith and colleagues (2006) suggest that officers repeated negative exposure to the connection between race, place, and crime can generate experientially based stereotypes that loosely connect civilian identities to their perceived criminality. Once these associations are solidified through the formation of cognitive schemas officers become more inclined to act on associations when identifying people, places, and situations that fit the mold for police intervention. This in turn can generate racial disparities in police behavior.

In addition to the development of cognitive schemas, Smith and Alpert (2007) contend that stereotypes and their influence on police behavior may be attributed to a separate albeit related theoretical process: illusory correlations. Rather than focusing on differential exposure and its capacity to develop stereotypes, illusory correlations rest on the assumption that officers may differentially retain information about peoples' identities and their association with crime. In drawing on the foundational work of Hamilton and Gilford (1976), Smith and Alpert (2007)

suggest that an illusory correlation exists when officers perceive a relationship between two things that are either not correlated, or correlated to a lesser degree than expected—such as one's race and their involvement in crime. Officers may unconsciously develop suspicion and overestimate criminality based on how individuals fit a particular stereotype, which in turn can generate oftentimes inefficient and racially-disparate police outcomes (Goel et al., 2016).

Taken together, racial disparities in police behavior may be the result of factors that create more opportunities for police to encounter non-White Americans. Although these factors generate racial disparities in police-citizen encounters, officers leading up to and during these encounters may not necessarily possess an intent to behave differently towards a racial group. Rather, the relationship between race and police outcomes is merely confounded by some objective factor that may explain the relationship once accounted for. For racial discrimination to exist in policing, differences in officers' *behavior* towards a racial group must be driven by an *intent* to treat that racial group differently. The following section outlines the proposed theoretical mechanisms that may give rise to discriminatory police behavior.

Theories of Racial Discrimination and Bias in Policing

Much like theories of crime, theoretical explanations for why racial discrimination exists in policing differ based on their assumptions about the extent and nature of the problem itself. Whereas macro-level theories assume racial discrimination is the product of cultural and structural factors that are external or internal to police agencies, micro-level theories assume discriminatory behavior is produced through internal or psychological factors among officers. The following section provides a brief overview of some commonly proposed theoretical mechanisms by their level of analysis.

Macro-Level Theories: Police Culture, Conflict, and Race

According to Crank (1998), racism in policing is produced by police culture. Like other occupational cultures, police culture provides officers with a toolkit to understand how to apply rules, principles, and information in their working environments (Manning, 2007). Police culture plays a critical role in the everyday functions of police officers because it gives them the collective knowledge of how to interpret emergent information on the job. Although scholars disagree as to what police culture consists of (see Herbert, 1998; Waddington, 1999), Paoline (2003) identified danger, suspicion, coercive authority, supervisor scrutiny, role ambiguity, maintaining the edge, laying low, crime-fighting, social isolation, and loyalty as consistent features across police culture conceptualizations. Several of these help explain how police culture can give rise to racial discrimination.

The working environment for most police officers is located within local neighborhoods and commercial areas (Korre et al., 2014). It is in these environments that the most salient cultural features of policing manifest—danger, suspicion, maintaining the edge, and coercive authority (Klinger, 1997; Muir, 1977; Skolnick, 1966). For example, while officers perceive a fairly low risk of being harmed during their shift, it is a consistently cited concern (Cullen et al., 1983). Some scholars believe this "danger paradox" is due to officers operating in uncertain environments where they are trained to maintain an edge to ensure their safety (Van Maanen, 1978). This overt preoccupation with the danger imperative—"a cultural frame that emphasizes violence and the need to provide for officer safety"—can have deleterious effects for people of color by way of shaping officers' attitudes and corresponding behavior towards them (Sierra-Arévalo, 2021, p.71).

For example, Skolnick (1966) describes how training officers teach police recruits what to look for when defining who is potentially dangerous or what he refers to as the "symbolic

assailant." Shared in the form of "war stories" and warnings, officers develop common sense knowledge about potential suspects that traditionally associate racial minorities with crime and violence (Rowe, 2023, p.136). Officers may interpret mannerisms such as the way people speak, walk, and appear as early warnings of their potential threat (Skolnick, 1966, p. 47-48). Likewise, Van Maanen (1978) argues that police may preemptively conceptualize racial minority groups as "outsiders" who are "…not to be trusted, are unpredictable, and are usually 'out-to-get-the-police'" (p.11). Collectively, officers draw from this knowledge base when associating observable actions with minority group members to identify suspiciousness and potentially discriminate against them.

As to where racial bias in police culture comes from, one explanation has dominated theoretical discourse. Scholars suggest that police cultures give rise to racial discrimination in policing when the overarching values, goals, and norms dominant in their institutional environment are premised on racial prejudice (Petrocelli, Piquero, & Smith, 2003; Stolzenberg, D'Alessio, & Eitle, 2002). This explanation operates on the institutionalist perspective of police culture, which suggests that police agencies are permeable institutions that transmit the dominant values, goals, and norms in society to uphold their legitimacy in that environment (Crank & Langworthy, 1992). Accordingly, minority group members will experience disparate police contact in part because the ruling White class perceived them as a threat to their cultural values, beliefs, and norms (Blalock, 1967; Holmes, 2000). This explanation is referred to as the racial/minority threat hypothesis and is frequently tested by examining the association between the relative size and geographic distribution of minority members in an area and police outcomes (Black, 1973; Liska, 1992).

Explanations of racial discrimination in policing through the racial threat hypothesis date back to the enforcement of antebellum slave codes and post-Civil War Jim Crow laws of the

American South (Hadden, 2003). These laws were used to typify Black Americans as part of the dangerous class and portray their resistance to bondage as a threat to be controlled for the preservation of the slave economy during the late 1600s and 1700s (Hadden, 2003; Reichel, 1988). As a result, slave patrols were created as a form of transitional police force, tasked with controlling the dangerous class by enforcing slave laws and incapacitating fleeing slaves (Reichel, 1988).[2]

Similar historical examples exist among critical analyses of the police during the Civil Rights Movement. In the late 1950s and early 1960s, Black Americans resisted dominant White, middle-class values and norms through peaceful protests that were often met with police force (NAS, 2018). In documenting police brutality during the Detroit civil rights protests in 1958, the NAACP reported that police frequently responded with elevated levels of force and racial epithets when Black Americans questioned their encounters with them. More specifically, over half of all complaints made by Black Americans to the NAACP included details about unprovoked assaults paired with racial epithets.[3]

Empirical tests of the racial threat hypothesis have yielded inconsistent results, due in part to the fact that they often fail to connect macro-level social processes to the development of police culture, and the subsequent manifestation of racial discrimination. Early empirical research was largely supportive, with many scholars pointing to racial composition in neighborhoods and cities as a determinant of police enforcement, lethal force, and complaints of misconduct (e.g., Holmes, 2000; Jacobs & O'Brien, 1998; Kent & Jacobs, 2005). In other words, larger minority populations were shown to be associated with racially-disparate police outcomes. Racial threat scholars

[2] It should be noted that these transitional police forces were not the foundation of American policing, but rather existed as a situational byproduct based on the need for formal social control in the South, which relied primarily on a slave-based economy. Policing in colonial America was structured by a formal sheriff, constable, and watch system that was originally heralded by England leading up to and throughout the seventeenth and eighteenth centuries (Monkkonen, 2004).

[3] https://policing.umhistorylabs.lsa.umich.edu/s/detroitunderfire/page/1958-63

interpreted such findings as evidence of discrimination—the ruling class (represented by the police) attempting to exert power over minorities. However, more recent empirical analyses challenge the robustness and magnitude of the associations between minority composition and police behaviors (e.g., Novak & Chamlin, 2012; Stolzenberg, D'Alessio, & Eitle, 2004; Stults & Baumer, 2007). Given these concerns and the challenges with measuring culture as the mediating link between racial threat and police discrimination, some scholars focus on individual-level explanations for racial discrimination to shed light on the more proximate indicators of such behavior.

Micro-Level Theories: Explicit Bias

Micro-level research directs attention toward the psychological processes that shape how individual officers perceive racial groups and behave in an intentional manner that conforms to their perceptions. In drawing from Smith and Alpert (2007), I define *intent* as broadly capturing officers' volition to treat individuals of a specific racial group a certain way. One micro-level explanation for racial discrimination involves the presence of explicit biases held among those officers who engage in such behavior.

Explicit bias can lead to racial discrimination through officers' conscious awareness of their racial prejudice of a specific racial group when choosing to treat them differently (Tomaskovic-Devey, Mason, & Zingraff, 2004).[4] Becker (1957, p.8) describes this generally as a form of "taste for discrimination" and may drive intent in a police officer's racially discriminatory behavior because of their racial animus towards them (NAS, 2018, p. 525). However, scholars contend that explicit biases, such as racial animus, are not strong predictors of racially biased police behavior (Smith & Alpert, 2007). One potential explanation is that general intolerance for

[4] Allport (1954) describes prejudice as being comprised of beliefs, attitudes, and behaviors. For this review, prejudice is narrowly referred to as attitudes, or what many refer to as racial animus.

racial prejudice has increased among the public and police agencies over the past two decades (NAS, 2018). Thus, some scholars posit that only a few officers may harbor explicit biases toward racial minority groups in an agency (Tomaskovic-Devey et al., 2004).

Another micro-level explanation rests on the assumption that although many officers may not harbor deep-rooted prejudices towards specific racial groups, race can still exist at the forefront of their decision to stop, search, and potentially arrest civilians. As described by Fagan and Davies (2000), civilians' race can be used as a criminal shorthand to determine suspicion by police officers. This has been reflected in many ethnographic accounts of how police consciously define suspicion (Muir, 1977; Rubinstein, 1978; Van Maanen, 1978). Smith and Alpert (2007) thus characterize the use of civilians' race under this theoretical framework of police discrimination as a source of information that is actively used in officers' decision-making processes. Officers are deemed "racial gamblers" who use civilians' race to determine potential criminality, and oftentimes inaccurately, stop and arrest those civilians (Smith & Alpert, 2007). The key difference between being a racial gambler and someone with an illusory correlation is perhaps unsurprisingly the same differentiating factor between disparity and discrimination: the presence of officers' intent to treat individuals of a specific race differently.

Summary

Taken together, theoretical explanations for the existence of racial discrimination differ in part by their level of analysis. Macro-level theories direct attention towards the social and cultural determinants of racial discrimination in policing. For example, police culture serves as an informational toolkit to help officers understand how to navigate the dangers and uncertainties of their working environment. Theorists posit that this can generate racial discrimination by socializing officers to look for situational cues or characteristics that inherently associate

criminality with racial minority status.

Micro-level theories of racial discrimination assume intent is the product of individual officer biases. While racial animus is declining in public polls (Saad, 2022), officers may intentionally and consciously treat people of a racial group differently due to their negative attitudes or beliefs about that group and their involvement in crime. This would correspond with recent survey findings by Roscigno and Preito-Hodge (2021), which concluded that police officers are "distinctly racist" when compared to the general public and other professions based on General Social Survey data from 1984-2018.[5]

Although this review treats police culture and individual officer biases as distinct explanations of racial discrimination, they may not operate in isolation. Police culture may provide a medium through which racial biases manifest in police behavior. Importantly, officers have individual agency and will exhibit variation in their adoption of culture rather than merely serve as passive conduits or "culture dopes" that perfectly reflect them (Paoline, 2003; Rowe, 2023). Officers bring their attitudes, beliefs, and ideas into the occupation as well, which helps them determine how they will enact "real policework" (Muir, 1977). Accordingly, police culture may initially operate at a monolithic, occupational level but it is filtered down through organizational contexts and within individual officers to fit their working environments (Campeau, 2015). This may help explain why those who contend agencies are presumably entrenched in discriminatory police cultures do not yield discrimination among all their officers. This may also help explain why such behavior can be concentrated in some police beats, precincts, or districts in the same agency.

Accordingly, while some attribute police culture as the primary source of discriminatory

[5] It should be noted that their estimates and substantive conclusions have been recently scrutinized and should thus be interpreted cautiously (see Peyton, 2021)

behavior (e.g., Crank, 1998), it may be through explicit biases that police culture eventually leads to racial discrimination. In other words, environments may give rise to discriminatory behavior due in part because of their ability to instigate and re-affirm biases towards a racial group of people among individual officers. It is then through individual officer biases that culture is indirectly associated with racial discrimination in a police agency.

To reiterate, racial discrimination in police behavior must involve intent. The causal mechanisms of racial discrimination are not necessarily the same as what generates disparate police behavior. Thus, disparity can serve as a potential signal of discrimination.[6] Accordingly, the next section provides an overview of the research that aims to identify racial discrimination and disparity in police behavior. In so doing, I highlight the challenges to estimation and their direct correspondence with the issue of discerning between racial disparity and racial discrimination in police behavior.

Research on Racial Discrimination in Policing

External Benchmark Tests

Early empirical tests of racial disparities and discrimination in policing were built on the assumption that analysts can estimate a population at risk for police contact (Neil & Winship, 2019). If so, analysts can construct a benchmark dataset from which they compare its racial distribution to that of a police outcome. These tests are referred to as "external benchmarks" because the reference data are external to the police agency. When a racial minority group is overrepresented in their distribution of a police outcome relative to their distribution in the

[6] There are, of course, situations where discrimination may be the result of racial differences that are not captured through data. For example, if police outcomes are systematically not reported for specific racial groups, this could be a product of discrimination. Yet, a measure of racial disparity in that outcome given the observed data may not capture the underlying discriminatory behavior—partly because it is not reported. Without available data on reporting biases, this discussion merely serves as a potential motivation for future research.

benchmark dataset, analysts conclude racial disparity and potential racial discrimination in policing.

For example, in Table 1, Black civilians comprise 67% of all arrests and 78% of all traffic stops in the hypothetical example, whereas they comprise 50% of the total population and 67% of those drivers involved in traffic crashes. Taking the difference in these percentage points yields a 17-percentage point difference, which indicates that Black civilians are overrepresented among arrests relative to their representation in the residential population. As demonstrated by Alpert, Smith, and Dunham (2004), the statistical significance of this difference can be estimated using a difference in proportions test, which involves dividing the difference in proportions by the standard error of the difference in proportions. In the case of arrests compared to the census benchmark, the estimated z-statistic (rounded to the nearest integer) for the difference in proportion of Black civilians in each dataset is 40, which according to standard levels of Type-I error for a two-tailed hypothesis test ($\alpha = 0.05$, $z = |1.96|$) would be deemed statistically significant. Put simply, Black civilians are over-represented among arrests given their representation in the census population.

Table 1. External Benchmark Demonstration

	Civilian Race		
	Black	White	Total
Police Data Source			
Arrests Population	67%	33%	15,000
Traffic Stops Population	78%	22%	1,275
Benchmark Data Source			
Census Population	50%	50%	200,000
Traffic Crashes Population	67%	33%	75,000
Percentage Point Difference			
Arrest % − Census %	17	-17	
Stop % − Crash %	11	-11	

The external benchmark test is the most popular and simplest method for examining racial disparities in police research, but it is not without its limitations (Smith et al., 2021). The main

assumption is that the benchmark dataset provides an accurate estimate of the population at risk for a police outcome (e.g., contact, arrest). Early research developed benchmarks based on census data, but these data do not accurately reflect the population at risk for such contact (Neil & Winship, 2019). Traffic collisions and observed driving population data were developed as promising benchmarks for traffic stops (Alpert, Smith, & Dunham, 2004; Lange, Johnson, & Voas, 2005), and arrest/offending benchmarks were created to reflect those at risk for police contact via pedestrian stops (Gelman et al., 2007).

Another major criticism is that these tests cannot distinguish between racial discrimination and disparity. For one, differences in a racial minority group's representation in a benchmark and their distribution in police arrests could be a function of differential offending, enforcement, or both—not simply officers' intent. This leads to a related criticism; benchmark tests are univariate tests that cannot account for external and situational factors that may contribute to a racial minority group's risk of being contacted by the police. As a result, external benchmark tests can only speak to racial disparity, they are usually interpreted cautiously, and they are often coupled with several other tests that will be described below.

Internal Benchmark Tests

Unlike external benchmarking, internal benchmark tests assume that most of a police agency's discrimination problems can be traced back to a few "bad apples" (Ba et al., 2021; Gonçalves & Mello, 2021).[7] Those interested in creating an internal benchmark thus assume that most officers in an agency are unbiased and can serve as a reliable reference group from which officers who discriminate can be distinguished (Ridgeway & MacDonald, 2009). In its simplest form, an internal benchmark is merely a dataset of stops, arrests, or other behavioral outcomes

[7] This assumption is derived from early empirical research and government reports that suggests most complaints and misconduct are tied to a few officers (Christopher, 1991; Sherman, 1978).

conducted by other officers in the agency. Analysts can compare the racial composition of those outcomes to that of an officer in question, and if Black civilians are over-represented in the officer's behavioral outcome, then they would flag them as engaging in racially-disparate police behavior.

These comparisons can be biased due to differences in what assignments officers work and where they work; thus, internal benchmarks can be improved by curating them for each police officer in an agency based on when, where, and what kinds of behaviors they conduct in their working environments.[8] The internal benchmark for an officer in question thus comprises a matched set of incidents of a pre-specified behavior (i.e., stops, arrests, uses of force) conducted by other officers in their agency during the same times, at the same places, and within the same contexts (Ridgeway & MacDonald, 2009; Walker, 2001). This creates a unique set of counterfactual data for each officer in an agency that allows analysts to determine which officers (if any) in an agency are the contributors of racial disparity.

To see how this works, Table 2 displays the distribution of a hypothetical officer's stops, and the distribution of their internal benchmark stops. This officer conducted a greater proportion of their stops at the beginning and end of the year, during the beginning and end of the week, and in precinct D. Their internal benchmark dataset comprises stops that have the same distribution on observed characteristics (see 3rd and 5th column). This is what creates the counterfactual necessary for the internal benchmark to be valid. The procedures used to estimate the internal benchmark varies, though modern empirical approaches leverage advanced matching procedures (e.g., propensity score analysis) to generate a weighted sample of stops for the benchmark dataset (Ridgeway & MacDonald, 2009; Walker, 2001). Therefore, the officer in question conducted 643

[8] There are cases when benchmarks may be created for officer workgroups, but these group-level benchmarks are not as common as the officer-level benchmarks.

stops while their internal benchmark is drawn from approximately 3,098 similarly situated stops based on time, place, and context.

Table 2. Internal Benchmark Demonstration

	Stops conducted by officer in question	% Stops conducted by officer in question	Internal benchmark stops	% Stops within internal benchmark
	N = 643		N = 3,098	
Month				
January	38	6%	183	6%
February	64	10%	308	10%
March	88	14%	424	14%
April	71	11%	342	11%
May	6	1%	29	1%
June	17	3%	82	3%
July	21	3%	101	3%
August	57	9%	275	9%
September	55	9%	265	9%
October	84	13%	405	13%
November	80	12%	385	12%
December	62	10%	299	10%
Day of Week				
Monday	127	20%	612	20%
Tuesday	57	9%	277	9%
Wednesday	61	10%	296	10%
Thursday	95	15%	457	15%
Friday	90	14%	432	14%
Saturday	99	15%	477	15%
Sunday	114	18%	547	18%
Precinct				
A	109	17%	526	17%
B	115	18%	555	18%
C	158	25%	760	25%
D	261	41%	1257	41%

The intuition follows that there is evidence of racial disparity if the racial distribution of an officer's stops differs from the racial distribution of stops conducted by their peers in the

benchmark dataset. The comparison effectively generates a *z*-statistic that estimates the significance of this difference in racial distributions and thereby solidifies analysts' conclusions about the existence of racially-disparate policing for an officer in question. Concerns over the potential false discovery of an officer as engaging in disparate behavior (i.e., a bad apple) have led to subsequent estimation procedures that account for this possibility as well by drawing on insights from the multiple hypothesis testing literature (Ridgeway & MacDonald, 2009).

It bears noting that the internal benchmark test has four major drawbacks. First, it cannot distinguish between racial disparities and racial discrimination. In its inception, the internal benchmarking test was designed as an early warning system, whereby agency supervisors could monitor officer behavior and identify statistical outliers that deviated from their peers (Walker, 2003). However, these deviations are often driven by factors that are indicative of racially-disparate policing rather than discrimination. For example, the TEAMS II Risk Management Information System used in the Los Angeles Police Department is criticized for flagging officers who, based on being among the top 1% of their peers, are deemed as statistical outliers despite potentially having legitimate reasons for their racially-disparate stop behavior. Spanish-speaking officers were frequently identified as outliers relative to their peers because they were often deployed to calls involving Spanish-speaking residents, which in turn, generated disparities in their contact with the Hispanic population (Ridgeway & MacDonald, 2009). This leads to a related issue with the internal benchmark approach. It can be difficult to create a customized benchmark for officers who rarely conduct stops and/or have highly specialized assignments or patrol locations. The issue is that some officers may have no peers who conduct similar stops as they do.

Another issue with the internal benchmark is that it requires extremely detailed data on officers' patrol assignments if analysts want to create matched benchmarks. For larger agencies

that collect high volumes of detailed data, with easily accessible records management systems, analysts can take advantage of this information to construct benchmarks that adequately reflect the reality of individual officers' routine patrol assignments. However, if analysts do not have access to this data, their benchmarks may not produce a valid comparison for each officer. Relatedly, if analysts do not measure specific variables that are associated with both the officer and the likelihood of stopping a non-White citizen, such as what neighborhood the officer works in, then the benchmark will be biased. Much like any other regression model built to explain social behavior, analysts will be constrained to the data available and the bounds of their theoretical and practical knowledge on the matter.

The last major limitation of the internal benchmarking approach is its inability to identify racial disparities if an agency is comprised of equally biased police officers. Put simply, if the entire agency is comprised of officers who discriminate, there is no way to individually identify any officer-in-question as being discriminatory because everyone behaves similarly. Accordingly, it may serve well to conduct tests of discrimination/disparity at an agency level prior to investigating any officer-level discrimination/disparity, as this will contextualize any null findings in the internal benchmark analyses.

Post-Stop Outcome Tests

Rather than analyzing police officers' decision to stop civilians, some analysts have instead evaluated their discretion in post-stop outcomes to identify racial discrimination. One approach involves the "outcome test," which has its roots in Becker's (1957) work to test for discrimination in labor markets. In police research, the outcome test operates under the assumption and null hypothesis that police officers are racially unbiased if the relative success in their decision to search individuals of a racial minority group is the same as they are for White individuals (Anwar & Fang,

2006; Engel, 2008; Knowles, Persico, & Todd, 2001). The typical outcome test involves measuring the proportion of "hits" or successful searches for contraband (e.g., illicit drugs, firearms) among all searches for a racial minority group compared to the rate for White civilians (Neil & Winship, 2019). Officers who are engaging in racially-disparate and potentially discriminatory behavior will be less (or more) successful in their searches of non-White individuals relative to Whites.

Advocates of the outcome test assert that observed differences in the hit rate *between* racial groups help distinguish between taste-based discrimination and statistical discrimination in policing (Anwar & Fang, 2006; Knowles et al., 2001). If policing is premised on statistical prediction, one would anticipate that, based on racial differences in serious offending, hit rates would be greater among non-White suspects compared to White suspects because officers are keen to enhance their efficiency in seizing contraband. While this assumes officers are discriminatory in their search behavior, such discrimination may be desirable given that it would indicate they are either more selective in their searches of non-White civilians or are simply more productive when choosing to search them. In contrast, taste-based discrimination exists when hit rates are lower for non-White civilians relative to White civilians. The idea of taste-based discrimination can more broadly be situated within Smith and Alpert's (2007) notion of the illusory correlation. When officers unconsciously over-estimate the association between criminal behavior and racial identity, they are exemplifying an illusory correlation, which in turn is reflected by their inefficient and fruitless encounters with nonwhite civilians relative to Black civilians. If, however, race was at the forefront of these officers' minds, then they would be Smith and Alpert's (2007) racial gamblers and unsuccessful ones at that.

Despite its potential benefits, the outcome test has two major flaws. First, little evidence

exists in support of a purely statistical prediction approach to policing.[9] Evidence suggests that specific proactive policing strategies such as the stop, question, and frisk program in New York City, were largely unsuccessful at seizing contraband despite disproportionately targeting non-White residents (Manski & Nagin, 2017). Likewise, the NAS (2018) contends that a pure efficiency model of policing is unconstitutional and does not conform to modern community-driven approaches to proactive policing.

Another flaw of the outcome test is the "infra-marginality" problem (see Table 3).[10] This occurs when the chances of a potential suspect exhibiting signals indicative of contraband vary by racial group and when the police are racially biased in their decision to stop civilians in the first place. To understand how this works, first assume that officers are trained to search for specific signals or cues indicative of contraband (e.g., illicit drugs, firearms), and these signals differ in their probability of indicating that the contraband is in someone's possession. This is displayed in the first two columns of Table 3: whereas 1% of people who carry contraband exhibit no signals of possession, 20% of people who carry contraband exhibit an odor indicative of possession (e.g., gun powder, marijuana smell).

Table 3. Inframarginality Problem Demonstration

	Probability of Successful Hit	% of Black Suspects Indicating Signal	% of White Suspects Indicating Signal	Black Probability of Hit	White Probability of Hit
Signal Type					
No Signal	0.01	70%	70%	0.01	0.01
Furtive Movements	0.10	15%	15%	0.02	0.02

[9] It should be noted that predictive policing—a place-based approach to proactive policing premised on forecasting when and where crime occurs in local areas—seemingly corresponds with statistical prediction as discrimination in policing. Despite recent criticisms about its implications for racial discrimination (Richardson, Schultz, & Crawford, 2019), preliminary evidence indicates that it does not generate racially biased policing (Brantingham, Valasik, & Mohler, 2018). Nevertheless, more research is needed using the outcome tests on predictive before any definitive conclusions can be made.

[10] See Neil and Winship (2019) and NAS (2018, box 7.1) from which this example and table is directly drawn from.

Table 3. (cont'd)

Odor	0.20	10%	15%	0.02	0.03
Both	0.50	5%	0%	0.03	0.00

Assume there are differences in the probability that a racial group will exhibit specific signals as well. Suppose that 70% of Black civilians who are stopped by the police exhibit no signals of possession, another 15% exhibit furtive movements, 10% exhibit an odor indicative of narcotics, and 5% exhibit both movements and odors. White civilians also exhibit similar signal distributions but 15% of those stopped exhibited an odor and none exhibited both furtive movements and odors. Randomly stopping and searching a Black civilian will thus yield a greater chance of discovering contraband compared to a White civilian. Taking the weighted average of these signal distributions for each racial group (see last two columns of Table 3) in this hypothetical example indicates that Black civilians are more likely to exhibit signals indicative of a higher probability of possessing contraband.

Now, consider that the police are racially biased in their decision to stop civilians. The police stop all Black civilians who exhibit furtive movements, odors, or both. In contrast, the police only stop all White civilians who exhibit an odor or both. In Table 4, 15% of all White civilians and 30% of all Black civilians will be stopped by the police. The implied hit rate for White civilians in an outcome test is .20, which is calculated by taking the weighted average of those stopped by the police for each of the signal groups that fall within the police threshold for stopping White civilians.

In this case, all White civilians in the sample of those stopped had exhibited an odor, thus two out of every 10 who exhibited an odor had contraband. The implied hit rate for Black civilians is also .20. However, the composition of those who exhibit signals among Black civilians differ

and the stop threshold for them is lower. Whereas 50% of Black civilians stopped by the police exhibited furtive movements, one-third exhibited an odor, and one-sixth exhibited both an odor and furtive movements. Calculating the hit rate across each signal group in the Black civilian sample that was stopped based on lower evidentiary standards is thus (50%(.1) + 33%(.20) + 17%(.50) = .20). Herein lies the infra-marginality problem: the hit rate for Black civilians is the same as for White civilians despite the racial bias that exists in officers' decision to stop Black civilians based on lower evidentiary standards. What this suggests more generally is that the outcome test will be invalid if the police are racially biased in their decision to stop people and there are differences in the racial distribution of signals indicative of contraband.[11]

Natural Experiments

Many of the tests for racial disparities and discrimination in policing mentioned above rely on observational methods that are limited in their ability to yield a valid estimate of disparity, let alone identify racial discrimination. Accordingly, some researchers have identified conditions under which there exists an exogenous source of variation in police discretion that can be measured to identify unbiased estimates of racial disparity and potential discrimination in policing.

One popular natural experimental method to test for racial disparities in police traffic stops is the veil-of-darkness hypothesis (Grogger & Ridgeway, 2006; Knode et al., 2024). The veil-of-darkness hypothesis assumes that police officers have less ability to visually identify drivers' race at night compared to during the day.[12] If so, race should not influence officers' decision to stop vehicles at night. Analysts can use stops conducted at night as a "race-blind" set of counterfactual

[11] Several solutions to the infra-marginality problem have been proposed over the years but their validity depends on data availability and the consistency of the underlying data generating processes for police stop and search decisions (Knowles, Persico, & Todd, 2001; Ridgeway, 2006).

[12] It is important to note that the difference in visibility between daylight and darkness is not assumed to be absolute. Rather it only assumes the difference in visibility is relatively greater in daylight than in darkness.

stops from which stops conducted in daylight—stops when race may more easily shape officers' decision to pull a driver over—can be compared.

For the veil-of-darkness hypothesis to produce a valid estimate of racial disparity, one must assume that the relative risk of being stopped by the police during periods of daylight and darkness is the same within each racial group. In other words, if daylight is the independent variable and drivers' race is the dependent variable, analysts must account for any confounding factors that relate to both variables. For example, differences in travel patterns over the course of the day may vary both within and between races, thus creating a differential risk of being stopped by the police in daylight and darkness. This will bias the veil-of-darkness hypothesis if unaccounted for. To account for these differences in risk of being pulled over during daylight and darkness, analysts constrict the sample of stops analyzed to only those that occur within the intertwilight period. This creates a sample of data that contains stops that took place in daylight while others took place in darkness at the same time of day, which allows for a direct comparison of those stops with otherwise similar circumstances.

In short, the veil-of-darkness approach leverages exogenous variation in visibility across daylight and darkness to determine if police officers are more likely to conduct traffic stops involving Black drivers during times of the day when drivers' race is visible compared to when it is not. Comparing the racial distribution of stops by officers in daylight and darkness therefore allows analysts to test for racial disparities in officers' decision to conduct traffic stops.

The veil-of-darkness hypothesis may not be able to identify racial discrimination in stop behavior under specific conditions. For one, the veil-of-darkness test uses a multivariable regression approach to estimate racial differences in officers' traffic stop behavior. This means that there may be legitimate reasons as to why officers pull over more Black drivers during daylight

as opposed to darkness that will bias the model results if left unaccounted for. For example, evidence suggests that failing to account for officer assignment and patrol unit type can bias estimates of racial disparity (Vito et al., 2020; Worden, McLean, & Wheeler, 2012). Moreover, the statistical power to identify racial disparities in the veil-of-darkness method is weakened when officers' decision to stop vehicles is influenced by car characteristics associated with racial groups (Grogger & Ridgeway, 2006).

Another natural experimental method proposed by Gonçalves and Mello (2021) leverages exogenous variation in officer discretionary ticketing practices to determine whether they racially discriminate when enforcing punishments for speeding. In their study of the Florida Highway Patrol (FHP), Gonçalves and Mello identify distinct "jumps" in the fine schedule for speeding that correspond with increasing levels of harshness (e.g., ticket fine increase, mark on driver's record). For example, driving 7-9 miles per hour over the legal limit results in a $125 fine whereas driving 25 miles over yields a $250 fine. The authors identified that officers have discretion in what speed to charge drivers and will typically "discount" the driver's speed to fall below the next jump in the fine schedule. In their study, they identified a notable jump in the fine schedule when shifting from 10 to 11 or more miles per hour over the legal limit. This corresponded with most discounting practices that involved listing drivers as speeding 10 miles over as opposed to their actual speed.

Akin to the veil-of-darkness design, Gonçalves and Mello (2021) construct a counterfactual group of "bias-free" officers who exhibited no discounting discretion in their ticketing practices. These officers and their stops were comparable in observed characteristics to officers who exhibited discretion in their ticketing practices. Gonçalves and Mello (2021) thereby use this natural experimental setting to determine if officers in the FHP were more likely to discount tickets for White drivers compared to non-White drivers. This ticketing discount design was robust to

racial differences in ticket discount requests made by drivers and driving locations. However, the primary limitation of their study is that the design cannot distinguish between racial disparity and discrimination.[13] A smaller though notable limitation to the study is that little empirical attention has been directed at the consequences of racial disparities in ticketing discount practices. Some evidence suggests that these behaviors may be linked to financial revenue-generating mechanisms in local municipalities, such as in Ferguson, Missouri (Department of Justice, 2015). Yet, far less is known as to how racial inequalities in ticketing practices relate to broader public safety and health consequences that are associated with racial inequalities in other police behaviors (i.e., arrests, uses of force).

Classical Experiments

An ideal approach to estimating racial disparities and potential discrimination in police behavior would be to construct an experimental design that randomly assigns officers into one of two conditions differing solely by civilian race: one condition involves a non-White civilian and the other involves a White civilian. Evaluating officers' decision-making between these conditions thus allows analysts to estimate differences in police behavior attributable solely to the race of a civilian with a high degree of internal validity.

Many racial discrimination experiments have been conducted in the field of cognitive psychology but the totality of evidence indicative of racial disparities and discrimination in policing is inconclusive. For example, in videogame-based shoot/don't shoot scenarios, some evidence suggests that officers have lower shooting thresholds and quicker reaction times when shooting Black suspects as compared to White suspects (Correll et al., 2002; Correll & Keesee, 2009; Taylor, 2020). Moreover, officer shooting error rates for shooting Black suspects are higher

[13] For similar natural experiments designed by economists to test for racial disparities and potential discrimination, see Fryer (2019), Weisburst (2019), and West (2019).

than for White suspects (Plant & Peruche, 2005). However, others have shown that officers can be quicker to shoot ("correctly") Black suspects and were more accurate in their decision to shoot Black suspects compared to Whites (Sadler et al., 2012). Moreover, some scholars have found that despite differences in reaction time to shoot by officers of civilians with different racial backgrounds, these same officers were no more likely to shoot Black suspects than White suspects (Correll et al., 2007).

A commonality among many experimental studies of shoot/don't scenarios is that they consist of videogame-based simulations, many of which involve research subjects pressing a single button to indicate whether they would shoot a potential crime suspect. As noted by James and colleagues (2016), these button-based simulations fail to capture the reality of a situation where officers may (or may not) deploy lethal force. The experimental settings lack the distinct environmental and situational factors that shape officers' decision-making when involved in police-citizen encounters, such as neighborhood context and what dispatch information officers are exposed to (Johnson, Cesario, & Pleskac, 2018). Moreover, they cannot capture the emotional dynamics and psychological toll that pervade officers' minds when they enter what Klinger (2004) refers to as the "kill zone" (p.14). Additional concerns of external validity with these simulations stem from the traditional use of a button to indicate whether officers will shoot a suspect. Pressing a button fails to replicate the experience of pulling a four to eight-pound trigger on a standard issue law-enforcement firearm, thereby leaving officers feeling as if they are actually in a video game rather than simulating the real-life experience of a shooting (James, Klinger, & Vila, 2014).

The lack of external validity with traditional shoot/don't shoot simulations has led to advancements in the design of these experiments. Recent research development has been directed at creating conditions premised on unfolding scenarios or what James et al. (2017) refer to as

"tactical social interaction" scenarios, which aim to simulate the many ways in which police-citizen encounters can escalate or deescalate based on officer input. However, research evidence to date using these more modern designs and simulators is limited and inconclusive (James et al., 2014; Johnson et al., 2018).

Summary

Despite the increased demand for transparency, researchers and practitioners alike are faced with a grim reality: there is no "silver bullet" to identifying racial discrimination. At best, each of these tests is designed to detect racial disparities in various police outcomes. At worst, these tests can provide misleading conclusions about the extent and nature of racial disparity and discrimination in policing (Neil & Winship, 2019).

Internal benchmarks and outcome tests are among the most externally valid and empirically robust approaches, yet they cannot explicitly identify discrimination. This is because they are built upon the research design principle of "closing back door" explanations that could account for disparate police behavior (Huntington-Klein, 2022). That is, they rely on methods that involve accounting for all observable factors that might explain such behavior with the caveat that any unmeasured factors could bias their findings.

Neither approach offers avenues to explore "front doors" through which credibly exogenous sources of variation in police discretion and intent can be measured to identify their unique causal effect on police behavior (Cunningham, 2021). The added benefit of these front-door approaches is that analysts can isolate variation in officer discretion to explicitly identify any potential racial discrimination in their behavior. Natural experiments and randomized control trials offer such an approach, but their external validity remains a critical issue that has yet to be resolved. These sentiments are best characterized by Neil and Winship (2019), who aptly stated

that "[t]he implication is that many of our tests of discrimination are not likely giving us the right answer, with there being more or less discrimination than results suggest" (p. 74).

Is testing for racial disparities in policing a fruitless endeavor? The public safety and health consequences of racially-disparate policing for people of color are readily apparent; thus, identifying when such behavior exists in an agency is a worthy endeavor for the prevention of its consequences. However, the question remains what should practitioners do when such disparities arise in a police agency? The next section unpacks some of the most commonly proposed police reforms and provides evidence of their effectiveness.

Policy Interventions for Reducing Racial Discrimination in Policing

A tragically predictable sequence of events follows the aftermath of high-profile police-involved killings of unarmed Black or African Americans. Whether it be Michael Brown in Ferguson (2014), George Floyd in Minneapolis (2020), or most recently Tyre Nichols in Memphis (2023), national dialogue on racial injustice and police reform rapidly follows their deaths. This is perhaps best captured by President Biden in his recent State of the Union address following the death of Tyre Nichols who said, "Let's commit ourselves to make the words of Tyre's mom true. 'Something good must come from this'" (Shear, Tankersley, & Kanno-Youngs, 2023). Yet, for something good to come, there must be consensus on how to achieve it.

There is considerable debate among scholars and the public as to how to prevent racial disparities and discrimination through police reform. In general, police reform efforts can be broadly organized into one of five focus areas: accountability, punishment, training, enforcement change, and recruitment/diversification/retention. Table 4 outlines each focus area, its implications for racial equality, and some popular examples. Given the breadth of reforms developed over the years, I provide only a brief overview of select reforms and highlight what evidence there is to

support them as they relate to remedying racial disparities in policing.

Table 4. Five Police Reform Focus Areas

Focus Area	Implication for Racial Disparity/Discrimination	Examples
Accountability	Enhance visibility and social control to prevent/deter racial discrimination/disparity.	Body-worn cameras Civilian oversight boards Duty to intervene
Punishment	Incapacitate problematic officers to prevent further harm.	Decertification Removing qualified immunity
Training	Train officers to handle situations better and thus avoid racial discrimination/disparity.	Implicit bias training De-escalation training
Enforcement Change	Reprioritize enforcement strategies to prevent racial discrimination/disparity.	Stop traffic and low-level crime enforcement Crisis-response teams
Recruitment, Diversification, and Retention	Enhance sworn-workforce diversity to increase sensitivity and reduce racial discrimination/disparity.	Gender representation Racial representation Education enhancement

Accountability

Accountability-based reforms build or enhance policies, structures, and data-generating mechanisms in police agencies to increase oversight of officers and deter them from engaging in problematic behavior. For example, body-worn cameras (BWC) increased in popularity throughout the U.S. not long after the death of Michael Brown in 2014. Some have attributed this proliferation to the fact that there was no video footage of the controversial shooting, which elevated public awareness about the shortcomings of pre-existing police oversight mechanisms

(Adams, 2021).[14]

As early as 2016, national data indicated that between 50-60% of police agencies across the U.S. were fully deploying BWCs. Advocates of BWCs contended that increasing police visibility will shed light on the discretionary behaviors engaged in by police officers, enhance officers' self-awareness, civilize potential suspects, and deter officers from engaging in questionable behaviors—especially racially-disparate uses of excessive force (Ariel et al., 2017; White & Malm, 2020). However, evidence of BWCs mitigating police misconduct and racial disparities is limited. In a recent Campbell systematic review, Lum and colleagues (2020) concluded that BWCs were not consistently associated with declines in officers' use of force and arrest activity. Moreover, evidence of BWCs reducing racial disparities in police contact and force outcomes is largely nonexistent. Part and parcel of these limited findings is the heterogeneity in officers' discretion to deploy BWCs, how camera footage is reviewed in police agencies, and to what extent they are used as a performance monitoring system (Adams, 2021).

Another popular accountability mechanism is a civilian oversight board. The President's Task Force on 21st Century Policing (2015) originally advocated for COBs as an external mechanism for reviewing sentinel events (e.g., high-profile police-involved killings), however, their broader purpose is to "…identify any administrative, supervisory, training, tactical, or policy issues that need to be addressed" in a police agency (p. 88). The intuition is that having an external oversight institution will ensure police agencies are meeting the public standards for procedural fairness and racial justice (CCJ, 2021a).

There are generally three types of COB models, each of which vary in form and function.

[14] Body-worn cameras have been used since the early 2000s in the United Kingdom and Australia, but their popularity in the states is much more recent (Goldsmith, 2010). Some point to the death of Trayvon Martin in Chicago of 2012 as an earlier though equally important flashpoint in the origination of body-worn cameras (Lum et al., 2020).

Investigation-focused COBs are designed to investigate complaints filed against the police and operate as a reactionary model for police accountability (DeAngelis, Rosenthal, & Buchner, 2016). Auditor-focused COBs assess the quality of completed complaint investigations by an agency, and review-focused COBs provide broad assessments of complaint investigations, training, and practice (DeAngelis et al., 2016).

Variations in the design and implementation of COBs have led to inconsistent evidence in support of their effectiveness and hindered scholars' attempts to identify what mechanisms within them generate positive police outcomes. While some studies report that COBs were associated with reductions in racial disparities (Ali & Pirog, 2019), others report associated increases in the rate of civilian complaints filed against the police of misconduct (Terrill & Ingram, 2016). Moreover, scholars contend that COBs will continue to be limited in their capacity to enact positive organizational change and prevent consequential police behavior so long as they cannot initiate disciplinary actions (Clarke, 2009; Hope, 2020).[15] For example, few COBs can compel officer testimony, and even fewer have the power to subpoena officers (Witkin, 2016). This lack of disciplinary and investigatory power coupled with a lack of funding, committee member expertise, and political support has led to concern over their effectiveness in enhancing police accountability altogether (Witkin, 2016).

Other accountability-based reforms focus on creating organizational policies to enhance transparency and address officer misconduct and discrimination. Duty-to-Intervene (DTI) policies provide officers with the power and security to intervene, interdict, and notify their superiors when their colleagues (even their superiors) engage in excessive use of force, discriminatory behavior, or general misconduct (see Jones-Brown et al., 2021). Advocates of peer intervention policies

[15] It should be noted that some civilian oversight boards have the authority to recommend disciplinary measures, though these are far and few between.

contend that police agencies can improve officer and civilian safety while reducing the chances of generating racial disparities in police behavior (Aronie & Lopez, 2017; Taniguchi et al., 2022). Preliminary empirical evidence supports this assumption, however, observed reductions in officer-involved deaths and racial disparities in the police force may be temporary (Dawson et al., 2022; Jones-Brown et al., 2021). One of the main critiques of peer-intervention programs is that they must be coupled with active bystander training for them to work (Taniguchi et al., 2022). Indeed, early evidence suggests that such training can increase officer buy-in to DTI policies, however, their downstream implications for the use of force and racial disparities have yet to be rigorously evaluated.

Punishment

Those adopting a punitive approach to police reform focus on identifying ways to remove problematic officers from the sworn police force and deter others from engaging in discriminatory behavior in the first place. Some have suggested enhancing the legal, financial, and occupational consequences associated with engaging in problematic behavior. One of the oldest punishment mechanisms in police agencies involves the dismissal of an officer based on engaging in disreputable, criminal, or immoral behavior (Kane & White, 2009). An early critique of dismissal/terminations is that it can be difficult to prevent terminated officers from being hired at other agencies in the same state (Goldman & Puro, 1987). That is, the problematic behavior engaged by an officer may be displaced but not eliminated. Accordingly, de-certification measures have increased in popularity over the years because they prevent the *intrastate* rehiring problem that occurs with some dismissals (Atherley & Hickman, 2013). Invalidating an officer's training credentials thereby makes them unemployable as a police officer in their commissioning state.

One of the primary limitations of de-certification policies is that they often cannot address

interstate rehiring issues. That is, they cannot restrict officers from simply earning their certification and becoming a police officer in an agency from a different state (Atherley & Hickman, 2013). Another limitation is that there exists heterogeneity in the standards for what is deemed eligible for de-certification between states, which in turn generates differences in its usage. For example, in their examination of de-certifications across states, Atherley and Hickman (2013) find that most (96%) Peace Officer Standards and Training (POST) agencies have the power to decertify officers based on felonious convictions, whereas just over two-thirds of reporting agencies decertify on the basis of failure to meet training or qualification requirements.

States also differ in the forms of punishment applied to de-certification cases, with some allowing officers to reapply sooner than others (if at all). Some have also noted a loophole in the de-certification process, whereby officers may resign while their de-certification investigation is pending, which allows them to effectively avoid being de-certified (CCJ, 2021b). Considering the historical implementation issues with de-certification across states, the International Association of Directors of Law Enforcement Standards (IADLES) developed the National Decertification Index (NDI). The goal of the NDI is to provide hiring police agencies with a national database of license revocations so that the issue of interstate rehiring is effectively eliminated.[16] A secondary goal to the NDI includes the standardization of de-certification reasons and their corresponding punishments. However, previous estimates of voluntary NDI usage indicate only 30 states report to the NDI and 22 use the NDI in their hiring procedures, which limits its potential impact (Atherley & Hickman, 2013). This has led to a dearth of information on its effectiveness for the reduction of police misconduct more broadly, and racial discrimination in policing in particular.

Another approach to punishment-based reform focuses on rolling back legal protections

[16] https://www.iadlest.org/our-services/ndi/about-ndi

outlined by the qualified immunity doctrine for police officers (and public employees more broadly) who are tried in civil cases. Qualified immunity protects police officers from civil and legal repercussions that stem from incidents where the constitutionality of their actions is questioned. As outlined in *Saucier v. Katz* (533 U.S. 194), a court must determine (1) whether the facts alleged or shown by the plaintiff make out a violation of a constitutional right, and (2) if so, whether that right was "clearly established" at the time of the defendant's alleged misconduct. The George Floyd Justice in Policing Act of 2021, which passed in the House but failed in the Senate, represents the most recent attempt by policymakers to roll back the scope of qualified immunity in police reform legislation.

Both *Campaign Zero* and the NAACP contend that qualified immunity can lead to the protection of racially discriminatory behavior by police officers that would otherwise be unconstitutional if not for the overly restrictive "clearly established" standards for determining the unconstitutionality of such actions.[17] In response to concerns about racial discrimination in policing, some states, such as California, Colorado, and Maryland, have begun restricting the scope of qualified immunity though many remain firm on protecting the doctrine. Moreover, many of these policy changes are recent and vary in depth, which means they are not yet amenable to empirical evaluation regarding their effectiveness in reducing racial discrimination in policing.

Training

Training-based reforms focus on enhancing officers' cultural sensitivity or mitigating their reliance on implicit/explicit biases. Training-focused police reforms have received considerable support in recent years with increased awareness of the importance of procedural fairness in mitigating the harmful effects of aggressive policing—especially against people of color (Nam,

[17] https://endqi.org/, https://www.naacpldf.org/qualified-immunity/

Wolfe, & Nix, 2022; Wolfe et al., 2020). Most recently, training-based reforms were advocated as part of the George Floyd Justice in Policing Act of 2021, which aptly stated the need for the Attorney General to establish "a training program for law enforcement officers to cover racial profiling, implicit bias, and procedural justice."

Implicit bias training is one popular program that attempts to reduce racial discrimination in policing. Such training programs were advocated heavily by the President's Task Force on 21st Century Policing (2015) and continue to be at the forefront of current police reform debates. Bias training programs attempt to change officers' attitudes and thinking patterns regarding how they interact with and treat people of color (CCJ, 2021c). The primary focus is to encourage self-introspection before, during, and after interactions with people of color to make sure officers are cognizant of their potential biases and how they may shape their behavior. Adding to this motivation is evidence that suggests officers' implicit biases may be dynamic and amenable to change through training programs (James, 2018).

Despite recent advocacy for bias training, these programs have been implemented across agencies for quite some time. According to a 2019 survey of 150 large police agencies in the U.S., more than 69% had completed some form of racial bias training (CBS, 2019). In one of the few evaluations of an implicit bias training program, Worden and colleagues (2020) found that the Fair and Impartial Policing training for the New York City Police Department had very little impact on officers' attitudes towards discrimination and their willingness to act without prejudice. They also found no evidence of any change in racial disparities in police outcomes following the program. More recently, a study of 3,700 police officers by Lai and Lisnek (2023) also found that bias training can increase officers' awareness of their biases following the rollout of an all-day diversity training program but its impact fades after one month after the training.

Another program type that has garnered considerable attention following expressed support from the President's Task Force on 21st Century Policing (2015) is de-escalation training (Engel, McManus, & Herold, 2020; Wolfe et al., 2020). Whereas traditional police training programs emphasize the importance of speeding-up the decision-making process by police officers when force may be deployed, de-escalation training focuses on slowing it down to convert otherwise volatile interactions into calmer ones (Engel et al., 2020). Advocates of de-escalation training programs claim that they can prevent officers from engaging in excessive use of force and may encourage them to utilize procedurally fair policing strategies, which can have positive implications for community trust and public safety (McLean et al., 2020). The Council on Criminal Justice (2021) also reports that this style of training may reduce racial disparities in police force outcomes.

Evidence in support of de-escalation training programs is sparse but promising. Early empirical analyses of training programs revealed that officers' attitudes towards procedurally fair communication improved after training but showed no evidence of their effectiveness on officers' behaviors (McLean et al., 2020; White et al., 2021). More recently, Engel and colleagues (2022) evaluated The Police Executive Research Forum's (PERF) Integrating Communications, Assessment, and Tactics (ICAT) de-escalation training in the Louisville Metro Police Department in 2019 and found several positive impacts on officer attitudes and behaviors. Officers reported being receptive to the ICAT training, they were more willing to utilize de-escalation tactics, and this corresponded with reductions in police use of force, officer injuries, and civilian injuries.

Limited evidence in support of de-escalation training may be the result of variations in how evaluators define and implement their training programs (Engel et al., 2020). Furthermore, the limited effects of de-escalation training on officer behavior may be the result of a pre-occupation

with training officers on *how* to de-escalate situations while failing to recognize that officers need training on *when* to de-escalate situations (Wolfe et al., 2020). Enhancing officers' capabilities to identify situational cues and respond through effective communication may provide the foundation necessary for de-escalation training to reap its intended effects on police behavior (McLean et al., 2020; Wolfe et al., 2019). However, more empirical evidence is needed before definitive conclusions can be made about its effectiveness in reducing excessive force and racial disparities in policing.

Enforcement Change

Some police reform efforts focus on shifting away from crime control and order maintenance models and toward public health approaches that emphasize trauma-informed strategies for policing. Traffic enforcement has garnered increased scrutiny due to evidence suggesting that traffic stops, and subsequent ticketing practices, can generate significant racial disparities (Pierson et al., 2020; Gonçalves & Mello, 2022) and be used as revenue-generating mechanisms for local municipalities (Department of Justice, 2015; Harris, Ash, & Fagan, 2020). Further fueling recent concerns about traffic enforcement are campaigns following the wake of Tyre Nichols, such as the "Stop the Stops" initiative from *Vera Institute of Justice*, which calls on policymakers, police agencies, and the public to support various traffic enforcement pullback reforms.[18]

Although enforcement-change advocates agree that there must be a change to traffic enforcement, there is disagreement as to what extent and how it should be changed. Most recently, attention has been directed towards eliminating the enforcement of low-level traffic violations (e.g., broken taillight). For example, in a recent *TIME* article following the death of Tyre Nichols,

[18] https://www.vera.org/ending-mass-incarceration/criminalization-racial-disparities/public-safety/redefining-public-safety-initiative/stop-the-stops

Johnson and Johnson (2023) advocated for police agencies to pull back their investigatory traffic stops and pretextual stops to minimize their disparate impact on people of color. Indeed, research by Roach and colleagues (2022) recently showed that Black drivers were more likely to be searched and less likely to possess contraband when analyzing several million traffic stops across multiple states in the U.S. They also found that investigatory stops amplified these racial disparities for Black drivers. Accordingly, scholars have argued that pretextual stops—traffic stops aimed at serving a purpose other than the official reason for the stop—are used by police officers to discriminate against people of color and have low returns on investment for public safety.

Complicating debates on traffic enforcement reform is that there is inconsistent evidence to support or oppose such pullback efforts. There is some evidence that suggests agency-wide pullbacks in traffic enforcement can negatively impact traffic safety. In a study by DeAngelo and Hansen (2014), the authors concluded that traffic injuries and fatalities significantly increased following a pullback in traffic enforcement due to mass layoffs in the Oregon State Patrol in 2003. A similarly negative effect of traffic enforcement on traffic collisions exists when examining increases in traffic citations and tickets. Using municipal budgetary shortfalls as an instrumental variable, Makowsky and Stratmann (2011) found that increasing the number of traffic tickets reduced the number of traffic collisions and injuries across municipalities in Massachusetts between 2001-2003. Most recently, Nix and colleagues (2024) found that large reductions in stops and drug arrests led to notable increases in violent and property crimes across neighborhoods in Denver following the aftermath of the COVID-19 lockdown order and the death of George Floyd.

However, other scholars report no negative impacts on traffic safety when agencies pullback in their enforcement efforts. For example, in their study of over 2,000 police departments between 2000-2018, Cho and colleagues (2021) found that traffic fatalities and crime remained

unchanged following significant pullbacks to traffic enforcement and arrest activity following officer deaths in an agency. Moreover, some have found that re-prioritizing traffic stops away from low-level infractions and towards more serious offenses may reduce racial disparities and discrimination in traffic stops and post-stop outcomes. In a study of traffic data from 2013-2016 in Fayetteville, NC, Fliss and colleagues (2020) reported that reprioritizing traffic enforcement towards crash prevention and away from crime control generated significant reductions in racial disparities and traffic crashes. Unfortunately, evidence on the impact of targeted pullbacks is quite limited; thus, the verdict is still out as to whether such reprioritization/pullback efforts are a valuable police reform approach.

Another style of enforcement change that has received increased interest in recent years is the use of co-responder models and trauma/crisis-response teams. As described by Lipsky (1980), police officers are faced with a variety of problems and expectations from their organization and the public. Police officers' continually expanding role in serving the public has led to an absorption of responsibilities they are not normally trained to handle (Thacher, 2022). Accordingly, the logic follows that agencies can reduce officer workloads and improve outcomes stemming from police-citizen interactions if they team-up with professionals who are more appropriately trained to respond to such situations.

Co-responder models work by pairing police officers with mental health service providers as a secondary response to an ongoing call for service, whereby the co-response team provides trauma-informed care to reduce civilian stress and improve access to mental health treatment following the incident (Puntis et al., 2018; Reuland, 2010). Unfortunately, evidence to date in support of co-responder models is limited, due in part to the novelty of the approach. However, as described by White and Weisburd (2018), qualitative evidence from co-response teams indicates

they have promise. They report that co-responder models can positively impact police-citizen interactions when deployed strategically in crime hot spots and when based on trauma-informed approaches.

A final approach that has garnered increased attention following a study of traffic ticketing discrimination by Gonçalves and Mello (2021), is to re-assign officers to low-risk settings where they are less able to discriminate in the first place. In drawing from rational choice-based theories of organizational misconduct and discrimination in labor markets (Arrow, 1963; Becker, 1968; Greve, Palmer, & Pozner, 2010), Gonçalves and Mello (2021) posit that re-assigning highway patrol officers in jurisdictions with lower concentrations of minority drivers can reduce the extent to which they discriminate in their traffic ticketing practices. This aligns with criminological applications of Clarke's (1995) theoretical framework on situational crime prevention, wherein reducing opportunities for offending through access control can help prevent such behavior from arising in the first place. Indeed, limiting access to prime victims for criminal offending has been shown to reduce crime (for review, see Felson & Boba, 2010). However, such policies have received little attention in the realm of racial discrimination in policing, though the preliminary evidence is encouraging.

Recruitment, Diversification, and Retention

Recent attention among police reform debates is centered on approaches that emphasize employing (and retaining) high-quality officers to enhance the quality of administering police services and to generate downstream benefits for public safety and racial justice. Police leadership is particularly receptive to these types of reforms. For example, when fielded questions about recruitment and retention during his campaign for Chief of Providence Police Department, Oscar Perez emphasized the importance of quality over quantity (Lavin, 2023). Similar sentiments were

reflected in qualitative comments among Illinois Association of Chiefs of Police members as part of a statewide survey administered by *Police One* (Wojcicki, 2022). In describing the challenges to recruitment, agency commenters mentioned issues surrounding the quality of candidates, with one mentioning "The quality is much poorer" than in years past. This not only reflects the challenges that agencies face when striving for quality over quantity but also their generally accepted desire to recruit and retain high-quality officers.

What characteristics define a "high-quality" police officer is open for debate, but some proponents suggest diversifying the workforce through female and racial/ethnic minority representation (Ba et al., 2021). For example, the *30x30 initiative* advocates for agencies to pledge they will have 30% of their full-time sworn police workforce as female by 2030.[19] Motivating this sentiment are two important bodies of evidence. First, recent national statistics show that full-time sworn female officers make up less than 15% of all police officers in local police departments as of 2018 (Gardner & Scott, 2022). This estimate varies by department size and personnel assignment, yet female representation remains low relative to their composition in the residential U.S. population. Full-time sworn non-White officers similarly make up a small proportion of the police force relative to their composition in the population (Gardner & Scott, 2022).

The second piece of evidence pertains to the empirically identified benefits associated with having a more diverse police workforce. Evidence from a meta-analysis showed that female officers were consistently less likely to use force during civilian encounters (Bolger, 2015), while other studies reported females as less likely to use excessive force when any force is required (Schuck & Rabe-Hemp, 2005) and generate lower complaint rates relative to their male colleagues (Gaub, 2020). A recent study by Ba et al. (2021) also found that female officers in the Chicago

[19] https://30x30initiative.org/

Police Department from 2012 to 2015 were less likely to use force against non-White civilians compared to their male colleagues despite conducting similar rates of stops involving non-White civilians. Scholars theorize that these positive associations with police outcomes may be due to female officers being better at using "soft" skills, thereby making them better communicators in the field and better equipped to de-escalate situations when they have the potential to become volatile (McCarthy, 2013; Schuck & Rabe-Hemp, 2005; Todak & James, 2018).

Evidence in support of racial diversification is less consistent. Informing racial diversity reforms are theories of symbolic representation and representative bureaucracy, which posit that non-White officers may be better at improving police-citizen relations by providing equitable policing because they can understand and address civilians' life situations when they are of a minority background (Theobald & Haider-Markel, 2009). More broadly, this suggests that police officers, as street-level bureaucrats, would be better at tending to the needs and interests of their social counterparts if they were to reflect their racial composition (Riccucci et al., 2014). In support of these hypotheses, Ba and colleagues (2021) reported that Black officers in the Chicago Police Department conducted fewer stops, arrests, and use of force involving non-White civilians than their White colleagues. Hoekstra and Sloan (2022) similarly found that White officers are more likely to escalate their levels of force in racially diverse neighborhoods when compared to non-White officers. Others have found that Black officers are less likely to provide discounts in their traffic ticketing practices when compared to White officers (Gonçalves & Mello, 2021).

However, some evidence suggests that Black officers may be more likely to arrest non-White civilians compared to White officers (Brown & Frank, 2006). More recently, Headley and Wright (2020) found that Black officers from the New Orleans police department were less likely to scale-up force against Black civilians but more likely to arrest non-White civilians in police-

citizen encounters. What may be contributing to inconsistent evidence is the fact that civilians are more likely to see some Black officers as no different than their White colleagues (Benton, 2020; Brunson & Gau, 2014). That is, hiring a more diverse set of full-time sworn police officers may be important for representative bureaucracy, but it fails to consider the social context in which police-citizen relations are situated (Brunson & Gau, 2014). Accordingly, it may be that diversification policies need other reform measures to enhance their potential for enabling change, potentially through accountability, training, or punishment mechanisms to ensure changes are organizationally focused and culturally motivated.

Summary

Police reforms that have been empirically tested, such as civilian oversight boards, implicit bias trainings, and crisis-response teams, vary in their area of focus but share an unfortunate reality. Most of these reforms do not have clear and consistent evidence to support their implementation as an effective way to reduce discrimination and disparities in policing. Those that remain untested share a related yet distinct reality: police agencies are more likely to adopt police reforms that others have already adopted, but they are far less likely to try something entirely new (Adams et al., 2022). Whereas the former issue is a matter of testing the continued existence of such reforms, the latter reveals a much grimmer outlook for empirical research and one that remains a difficult challenge for current police reform. How then do researchers test police reforms if agencies do not formally implement them?

Fortunately, there are ways to explore questions about police reforms, such as diversification policies, without finding the natural experimental setting where a situation conducive to evaluation exists. For example, recent empirical research utilizing econometric-informed techniques has conducted policy simulations that depict the implementation of a policy

without actually doing so (e.g., Chalfin & Kaplan, 2021; Gonçalves & Mello, 2021; Hoekstra & Sloan, 2023). Their results are informative, replicable, and rooted in causal inference. Despite their potential for informing current police reform debates, few have conducted such analyses as they relate to reducing racial disparities in policing. This is concerning given the increasing demand for evidence-based policing and calls for data-driven police reform (Ridgeway, 2018; Sherman, 2013). In the next section, I draw on these insights to describe the foundation for my study and its implications for the advancement of evidence-based research on racial disparities in policing.

Current Study

At the beginning of this chapter, I reviewed the conceptual challenges that researchers face when investigating the potential presence of racial discrimination and disparities in policing. Complicating matters is that these unique concepts have distinct causal mechanisms, which are difficult to empirically distinguish—let alone identify—with leading research methods. Despite these differences, racial discrimination and disparities in policing share a host of indirect public health and safety consequences for people of color.

National dialogue on police reform has contributed to several solutions to racial disparities in policing such as enhancing police accountability, training efforts, and even diversifying the sworn workforce. However, very few of these reforms have been extensively and rigorously evaluated. Making matters worse, some agencies may be apprehensive about adopting such reforms due to a lack of adoption elsewhere (Adams et al., 2022), which may impede ongoing efforts to enhance racial justice.

For the etiology of racial disparities and discrimination in policing to grow and instigate positive change, there needs to be a push for more research that assesses the effectiveness of policy interventions on racial disparities rather than primarily identifying them. This is not to suggest

scholars disregard the importance of studying how to identify racial inequalities in policing. However, scholars must begin to consider *identification* and *remediation* in the same breath when studying racial disparities in policing. This will go a long way in connecting the dots between where the state of the research is and where it will be in the future.

The current study provides a first step in that direction. Before doing so, I outline the theoretical motivations and empirically backed justifications for why I used my identification strategy for racial disparities in policing, and why I chose to assess specific policy questions upon identifying such disparities. Providing a transparent connection between *identification* and *remediation* is critical from a methodological perspective because it ensures that they operate at the same levels of analysis when responding to racial disparities in police behavior. Put simply, I need to make sure that the solution fits the scope of the problem. It also highlights the importance of theory and its implications for policy and practice by providing context on why I would expect specific identification strategies to inform different reform strategies.

Knowing which identification strategies work best for specific policy interventions also has practical relevance. Academics bear the important responsibility of translating their research into practice (Lum & Koper, 2017). Providing policy makers and police leaders with a direct connection between how to identify disparities in police agencies, and what interventions to use based on that process and level of analysis is critical to successful reform efforts (Lum, 2009).

This study aims to build upon ongoing research on racial inequalities in policing by providing evidence of its existence in the Chicago Police Department (CPD) from 2012-2015, and to what extent it is associated with factors related to diversification reforms and officer re-assignment initiatives. The first step of the analysis, *identification*, involves investigating the potential presence of racial disparities in the CPD at an officer-level. Unlike past research, which

has conducted tests of disparity at an aggregate level, this study seeks to identify which officers (if any) engaged in racially-disparate police behavior. Motivating the use of the internal benchmark procedure is the fact that police officer misconduct can be concentrated among a few officers in the agency (including CPD) (Chalfin & Kaplan, 2021; Christopher, 1991), and this can be true for racial disparities in traffic stop behavior as well (Ridgeway & MacDonald, 2009). More generally, this stems from a rich history of criminological research that has shown that a small number of offenders produce a large proportion of crimes reported to the police (Wolfgang, Figlio, & Selling, 1972). Accordingly, if the problem is concentrated among a few officers, the internal benchmark approach offers the best chance of uncovering the problem.

Using the internal benchmarking approach to identify disparities in the CPD offers a natural segue into what kinds of policy interventions can potentially address the problem as well. The internal benchmark approach will provide a list of officers that conducted a significantly high (or low) proportion of stops involving non-White civilians relative to their peers in the same times, places, and contexts. Accordingly, policy interventions premised on diversification, re-assignment, or training must be applied to individual officers given that disparities were identified at that level of analysis. And while most interventions can be tailored to at-risk officers, there are a set of costs associated with singling officers out for intervention. These costs range from financial burdens associated with training programs to legal and climate concerns associated with punishing or enhancing the supervision of specific officers.

Scholars must, therefore, consider evaluating those policies that stand the best chance of being adopted given the scope of the problem that was determined by the method of identification. Arguably the most desirable officer-level interventions are those that incur either no costs to an agency's officers or only induce costs directly to the agency. For example, diversification

interventions have no direct impact on those officers who were identified as problematic but will incur direct financial costs to the agency through recruiting, hiring, and retaining a more diverse set of officers. Similarly, changes to patrol and shift assignments may inconvenience officers but they do not incur a financial burden to the agency.

Each policy has ample theoretical foundations to support hypothesis testing, which is critical to understanding why such policies may succeed (or fail) in the following analyses. Diversification policies draw on theories rooted in representative bureaucracy, thus hypothesizing that having a more diverse and representative workforce will reap tangible reductions in racially-disparate police behavior. Similarly, re-assignment policies rooted in situational crime prevention and rational misconduct in organizations would hypothesize that officers' racially-disparate behavior should become less prevalent in contexts where there are fewer opportunities to perpetuate such behavior. Accordingly, these stand the best chance of being deployed following the internal benchmark analysis given their potential costs and theoretical basis.

In the second step of the analysis, I draw from theoretical and empirical research on racial and gender diversification in the sworn police workforce to determine whether indirect evidence exists to support the hypothesis that such reforms can reduce racial disparities in policing by promoting more racially equitable behavior. Both gender and racial diversification policies assume that enhancing agency representation can improve police-citizen relations because officers can better tend to the needs and interests of their social counterparts if they more closely reflect their demographic composition (Riccucci, Van Ryzin, & Jackson, 2018). Gender diversification theorists further posit that female police officers may be better communicators and thus reap downstream implications when they use force, particularly against people of color (e.g., Ba et al., 2021). Using this information, I construct a dataset of at-risk officers who were previously found

to have engaged in racially disparate behavior. I compare behaviors engaged by these at-risk officers to subsets of officers who were not found to engage in disparate behavior by their race and gender. After making these comparisons, I explore whether female and Black officers have less racially-disparate police behavior than their racial/gender counterparts.

I then explore an alternative approach to addressing racial disparities in policing by assessing whether officers' disparate behavior can be modified based on the racial composition of the police beat they are assigned to work. Officer reassignment policies build on situational crime prevention propositions to suggest that placing officers in low-risk settings can help agencies potentially mitigate their racially-disparate police behavior. More generally, they assume that officer behavior is influenced by the racial composition of where they work. Accordingly, I test this assumption to see whether there is any indirect evidence in support of this. In the next chapter, I discuss the data analyzed in this study, the research design used to frame the analyses, and the empirical procedures used to carry out these analyses.

CHAPTER 3: DATA AND METHODS

Background

Research on racial inequalities in policing has recently shifted attention to more micro levels of analysis to understand how such behavior manifests among individual officers and their peers. This shift corresponds with a growing body of evidence that suggests racially-disparate behavior may be concentrated among a few "bad apples" in a police agency (Goncalves & Mello, 2021; Ridgeway & MacDonald, 2009). While this lends support to a widely contested idea surrounding racial inequalities in policing (Abraham, 2023; French, 2023; Shoub, 2023), few methods in this area of research are readily capable of testing the validity of the argument.

The internal benchmarking approach offers researchers one method to probe racial disparities at an officer level; however, it is not without its limitations. Some limitations are related to methodological shortcomings with its design while others are related to data used in the approach. To better understand these limitations, consider how one might conduct an internal benchmark analysis in an ideal research setting. Analysts would compare the racial distribution of stops involving non-White civilians by an officer in question with the racial distribution of stops conducted by other officers in the agency. Stops conducted by the officer in question comprise the treatment condition, and all stops made by their peers comprise the untreated condition. If each stop had an equal probability of being conducted by the officer in question and their peers, the treatment effect is identified through a binary response model of the form:

$$\Pr(Nonwhite = 1|t) = \beta_0 + \beta_1 Treatement + \varepsilon \qquad \text{Equation (1)}$$

where β_1 estimates whether there are significant differences in the probability of a stop involving a non-White civilian between an officer in question and their peers. Here, a "bad apple" constitutes an officer in question whose probability of conducting a stop that involves a non-White person is

significantly different from their peers based on a standard two-tail null hypothesis test.

Unfortunately, the basic internal benchmark analysis operates on a heroically unrealistic assumption about policing: civilians at any given time and place are equally likely to be stopped by the officer in question as they are to be stopped by other officers in the agency. In other words, the chances of being stopped by the treated officer or their peers are as good as random. The reality is that police officers in an agency will have different assignments and work different shifts within them, which makes it difficult to compare officers on a one-to-one basis. Within these shifts and assignments, officers may also exhibit micro-level variation in where and when they choose to interact with civilians.

More broadly, this issue of comparability indicates an important causality problem with the basic internal benchmark analysis: there is a difference in the propensity to be treated, and this difference may be due to factors that are jointly related to the probability of a stop involving a non-White citizen in the first place. When these confounding factors are unaccounted for, the internal benchmark analysis will yield a biased estimate of β_1.

Fortunately, Ridgeway and MacDonald (2009) proposed a matching procedure that accounts for the unique variation in each officer's stop behavior and the potential confounders related to the race of a civilian in a stop to improve the internal benchmark analysis. This broadly involves matching an officer's stops to stops conducted by their peers based on when, where, and under what contexts they take place to relax concerns of comparability and induce an evaluation based on common circumstances. However, this matching procedure involves incorporating extremely detailed data that most police agencies are not readily capable of providing, such as information on officer demographics, shift assignments, misconduct records, and enforcement activity. Data accessibility issues may be due to challenges with obtaining and manipulating data

in their record management systems, legal resistance to providing any identifying information about their employees, or more political reasons. However, when such data are readily available, Ridgeway and MacDonald (2009) make a compelling case for why and how their improved internal benchmark analysis can help agencies determine whether any of their officers are responsible for racially-disparate behavior.

The current study seeks to build on Ridgeway and MacDonald's (2009) work by utilizing a rich source of data that comes from the Chicago Police Department (CPD) to provide an empirically robust internal benchmark analysis and assessment of policy-relevant questions. These data are advantageous for three reasons. First, multi-behavioral assessments are needed yet lacking in current research on racial disparities in policing (Neil & Winship, 2019). This study builds on past research by measuring racial disparity through two common forms of police behaviors: stops and arrests. Findings from this multi-behavior assessment will provide a more concrete understanding of the extent and nature of racial disparities in the agency.

Second, these data provide an exceptionally high level of detail about officer activity, which supports robust officer-to-officer comparisons when incorporated into the internal benchmark analysis. Recall that Ridgeway and MacDonald's (2009) primary advancement to the internal benchmark analysis involves creating customized internal benchmarks that match each officer's stops based on when, where, and in what context they take place. Fortunately, data for this study includes information on where officers chose to engage civilians and important shift assignment information. Information about their patrol assignments, such as what shift they were assigned to in a beat, and when their stops and arrests took place were also readily available in the data for this study. This information helps maximize the improvements to the internal benchmark analysis proposed by Ridgeway and MacDonald (2009) to provide a robust racial disparity

analysis.

Apart from leveraging the advantages of highly detailed data and an improved methodological design, the internal benchmarking approach boasts high external validity. Unlike natural or classical experiments, which can be set in unrealistic or rare settings (e.g., shoot/don't shoot experiments) and/or analyze uncommon outcomes/behavior (e.g., ticket discounts, shoot button pressing), internal benchmarks analyze routine police behaviors in real-world settings with readily apparent consequences. In addition, growing empirical evidence suggests that a nontrivial proportion of an agency's misconduct, disparity, and force problems may be concentrated among a few officers (Chalfin & Kaplan, 2021; Sherman, 1978; Walker et al., 2001); thus, marrying an empirical approach to this reality is both realistic and theoretically relevant.

The next section provides an overview of the data analyzed in this study. Here, I describe three versions of the data. All raw datasets coming from their original source are referred to as "original datasets." These original datasets are not publicly accessible; however, they were pre-processed by a team of researchers and their data collection team to construct a series of "general use datasets" that are publicly available. All data sampled in this study come from these general-use datasets and comprise the "analyzable datasets." In the following section, I begin by describing the data sources from which the original datasets were created. In so doing, I explain how the original data were collected and how I obtained the general-use datasets. I then discuss how the original datasets were merged by the primary data collection team and structured as general-use datasets, paying particular attention to their levels of analysis. I also discuss how I cleaned the general-use datasets and merged them to construct the analyzable datasets for the analyses in this study.

Original Dataset Sources

The general-use datasets were prepared by Ba and colleagues (2021). I gained access to these datasets personally through permission from Dean Knox and Jonathan Mumolo, who were responsible for data collection in collaboration with the *Invisible Institute* and *Lucy Parsons Lab*. These organizations are registered 501(c)3 non-profits that seek to enhance police accountability by providing publicly accessible data on various police behaviors and law enforcement surveillance programs in the city of Chicago.

As part of a data sharing agreement, all data analyzed in this study come from the general use datasets, which are publicly available through a data repository, CodeOcean, which is hosted by the Research on Policing Reform and Accountability lab.[20] As such, the analyzable datasets in this study do not represent the original datasets as they were passed through a cleaning and merging process by Ba and colleagues before becoming publicly accessible. More details on this process will be described shortly.

The original datasets span five sources. Original data containing officer demographic information come from: (1) rosters of all available current and past Chicago police officers up to 2018; and (2) unit history data for individual officers from 1930 to 2016. These data were released following a series of open-records requests to the Chicago Police Department (CPD), Chicago Department of Human Resources, Chicago Office of Police Accountability, and the Illinois Office of the Attorney General. The requests were made by a team of researchers and the *Invisible Institute*, which were fulfilled after civil litigation (*Kalven v. City of Chicago*, 2014). The case held that documents related to allegations of police misconduct are public information in the state of Illinois. This led to a watershed movement regarding data availability on Chicago police officers

[20] https://policingresearch.org/

as well. Studies analyzing these data have featured in prominent outlets across disciplinary fields and bear important implications for the study of policing (Chalfin & Kaplan, 2021; McCarthy et al., 2020; Rozema & Schanzenbach, 2019).

The original dataset on officer shift assignments comes from CPD's automated Daily Attendance and Assignment sheet for each district spanning 2012-2015 while an original arrest dataset comes from CPD's internal data; each dataset was produced through a series of FOIA requests made by the data collection team. The original stops dataset was provided by the Lucy Parson's Lab, which collected "Stop, Question and Frisk" data through a series of FOIA requests to the CPD between 2012 and 2015.

General Use Dataset Preparation

In preparation for their study, Ba and colleagues (2021) merged the original datasets to create what they refer to as "activity profiles" for officers in the CPD. The goal of their merging procedure was to identify the demographic makeup of an individual officer and track their behavior (i.e., stops, arrests, uses of force) over four years based on what patrol assignments they worked from 2012 to 2015. This involved merging the roster and unit history datasets for individual officers to the assignment data, which were then merged with the stop and arrest data. As described by Ba et al (2021), each file was merged based on identifying information such as an officer's birth year, race, and gender or other characteristics (name, badge, appointment date, current unit). Before merging, this information was used to de-duplicate data within each dataset based on inter-file unique identifiers. The merging process followed by having officers in each dataset repeatedly merged on the identifying characteristics, whereupon any successful one-to-one match was then removed from the next merge attempt. The original datasets were thus consolidated into four general-use datasets (i.e., officers, assignments, stops, and arrests) that are linked through a unique

officer identifier that they created after the merging process.

General-Use Dataset Structure

Each general-use dataset created by Ba and colleagues (2021) has its level of analysis corresponding to a logical observation in the dataset. For example, the "officers dataset" is measured at the officer level and contains information on each officer's race, appointment date, resignation date, gender, and Spanish-speaking ability. The "assignment dataset" is measured on an assignment-by-assignment basis, with each row corresponding to a unique assignment. Each assignment contains information such as what beat, shift, and date the assignment corresponds to.

The "stops dataset" is measured at the stop level, meaning that each stop incident has at least one stop per officer involved in the incident. Chicago patrols are commonly conducted in officer pairs; thus, stop incidents frequently involve more than one stop observation. According to the general use dataset on stops, over 79% (N = 756,246) of the 946,912 stop incidents had two officers involved, whereas the remainder involved a sole officer.[21] The stop data includes information on who was the leading officer in a stop, where the stop took place (latitude, longitude), what form of contact the stop took (traffic, pedestrian) when it took place, and what the race and gender of the stopped civilian was.

In contrast, the arrest data do not indicate the primary arresting officer, however, over 95% (N = 157,067) of the 164,802 arrest incidents involved two officers according to the general use dataset on arrests.[22] Arrest data include information on what type of crime led to the arrest (e.g., violent, property, drug), the statute listed for the arrest (e.g., warrant, domestic battery), where and

[21] Recall that these data come from "Stop, Question and Frisk" data provided by CPD. Therefore, it represents a subset of all stops officers made in the CPD during this time. Whether other forms of stops outside this dataset involved more than 2 officers is unknown but certainly possible.

[22] The remaining 4.7% of arrest incidents (N = 37,876) were conducted by a single officer, and one incident was conducted by four officers.

when the arrest took place, and what the gender and race of the arrested civilian were. The following section discusses how these data were filtered and prepared to create the analyzable datasets in this study.

Analyzable Dataset Preparation

I merged the general-use datasets according to their unique officer identifiers to create two analyzable datasets for the internal benchmark analysis, one for each set of officer behaviors: a stop-level dataset and an arrest-level dataset. Each logical observation in these datasets corresponds to a stop or arrest made by an officer in the CPD, which contains information on when (date of incident, time of incident), where (latitude, longitude), and under what circumstances these behaviors occurred (beat assignment, stop type). Multiple sample restriction criteria were incorporated into this process to prepare the general-use datasets as the analyzable datasets for this study.

Beat Patrol Assignments

One of the unique features of CPD's patrol assignment process is that officers are designated to both standard and nonstandard beats for an assignment. As seen in Table 6, roughly half of all assignments involve officers patrolling in a standard beat that corresponds to a recognized administrative boundary on CPD maps. For example, officers are regularly assigned to beat number 0122, which corresponds to CPD's district 01, sector 02, and beat 02. This spans roughly 0.75 square kilometers, is in the heart of Chicago's business district, and covers some of its most famous buildings including the Willis Tower, Rookery Building, and Chicago Board of Trade Building.

Within this beat location, officers may be given either a regular patrol task "0122" or a relief patrol task with a designated suffix "0122R." For a complete breakdown of these

assignments, see Table 5. Relief assignments are for officers working the first shift (often referred to as "First Watch" internally) in this beat whereas second and third shifts are designated as standard patrol assignments. In both cases, these assignments were included in the data to ensure officers are fairly compared against one another based on where and what shift they work. Beat codes for both standard and non-standard assignments can also include an assigned suffix from the phonetic alphabet (i.e., A, B, C, D) for officers who are working in a squad that rotates based on the CPD operations calendar.

Table 5. Patrol Assignments by Beat Location

Beat Location	N	%
Standard Location		
Regular Patrol	976,293	27.70%
Relief Patrol	827,668	23.50%
Nonstandard Location		
Regular Patrol	2,245,927	63.80%
Desk Duty	297,298	8.40%
Total	3,519,518	

Patrol assignments can also designate officers to nonstandard beat locations, where no such official documentation exists of their exact whereabouts. As described by Ba and colleagues (2021), nonstandard beat locations may be drawn for administrative or community-based needs such as community meetings. These nonstandard beat locations also receive patrol assignments on a similarly frequent basis, thereby representing a unique set of places where CPD officers routinely interact with residents. It is for this reason that assignment data in these nonstandard beat locations were included in the analyses. A small proportion of assignment data also contain "desk duty" assignments that correspond to a nonstandard beat location with a suffix "02" (e.g., 1402). These assignments were omitted from the analysis as they did not pertain to true patrol assignments and made up a small proportion of all assignments (N = 297,298, 8.4%). Collectively this beat assignment information was used for operationalizing assignment information, which will be

described in more detail shortly.

Another unique feature of the assignment general use dataset is that there is a field indicating what rank the officer is for each assignment. This is particularly valuable as the internal benchmark analysis is premised on comparing officers who engage civilians based on common circumstances, one of which is their rank. Accordingly, all 2,809,920 (87.6%) of the 3,261,698 assignments designated for a "Police Officer" were retained in the analyzable dataset while the remaining 451,778 assignments were dropped to enhance comparability. Of these dropped assignments, the majority (61.1%, N = 275,869) were designated as "Sergeant" assignments while the remaining were slated for special tasks or ranks (e.g., Chief, Detective, Field Training Officer, Commander, Helicopter).

Stops and Arrests

According to the general use datasets, there were 946,912 unique stop incidents and 164,802 unique arrest incidents, many of which had two officers in each incident. These data were cleaned such that there is only one officer tied to each stop/arrest to ensure the internal benchmarking analyses do not compare stops/arrests made by an officer in question to the exact same stops/arrests that were made by their fellow officers involved in the same incident.

For the stop data, this simply involved retaining only the officer whom the data had designated as the primary stopping officer. This led to a reduced sample of 940,693 stops. Information on which officer was the arresting officer was not available in the arrest data. For each arrest, I randomly assigned one of the officers involved in an arrest as the primary officer who made the arrest. Randomly assigning officers in this way ensures I do not introduce any systematic bias into the data by safely assuming that an officer's decision to arrest a civilian was shared equally between them and their partner(s). This led to a reduced sample of 164,801 arrests. Upon

merging the officer assignment information that corresponds to the officer who made a given stop/arrest, the sample of stops was reduced to 789,901 stops and 132,520 arrests. These reductions were attributable to missing assignment information for the officer on the day on which they made a given stop/arrest.

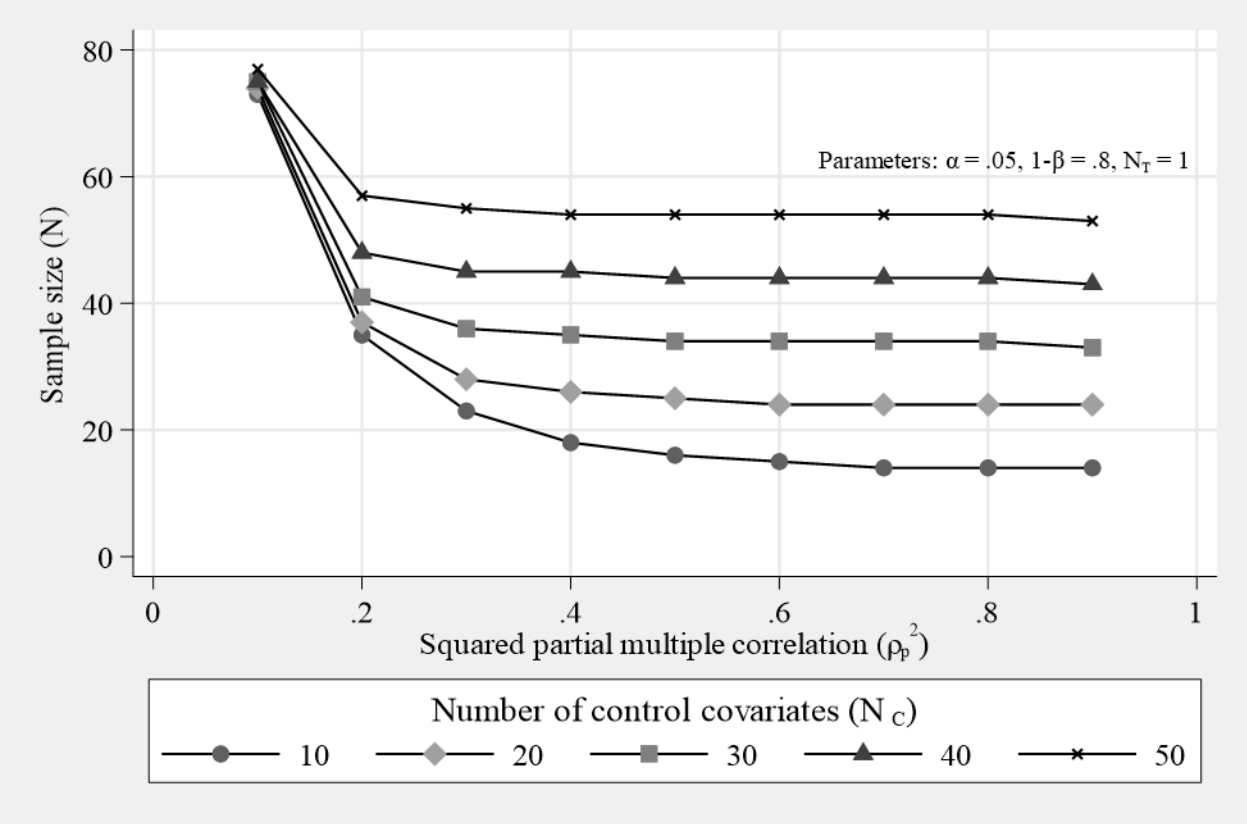
In preparation for the internal benchmark analysis, three sample restriction criteria were applied to the stop and arrest data. First, all stops and arrests conducted by personnel that were not deemed "Police Officer" were removed from the analyzable dataset of stops and arrests. This was a restriction criterion that carried over from the assignment dataset. Second, all unique incidents involving limited officer discretion when choosing to engage a civilian were omitted from the analyzable datasets. In the stops dataset, this included any stop incident resulting from the CPD's Repeat Offender Geographic Urban Enforcement Strategy (ROGUES) (0.01%, N = 68).[23] For the general use dataset on arrests, this included omitting any arrest incident stemming from a warrant (11.86%, 15,727), failure to appear in court (0.49%, N = 646), and out-of-state warrant (0.33%, N = 436).

Lastly, officers who conducted less than 50 stops or arrests over the course of the entire study period were omitted from the analyzable dataset of officers. This was informed by Ridgeway and MacDonald (2009), who stated that such a restriction is necessary to reach the bare minimum statistical power required for the internal benchmark analysis to follow. A power analysis was conducted using Stata 18 to further determine the risk of low statistical power. As displayed in Figure 1, the requisite number of stops or arrests increases as the number of covariates included in the internal benchmark analysis increases. The number of covariates will not exceed 50 given the variables included in the analyses conducted for this study, though this may vary in other study

[23] Interestingly, no publicly available documentation exists on this program. As such, it was omitted from the analyzable dataset.

settings. In contrast, the requisite sample size decreases as the anticipated effect size increases. Assuming a standardized effect size for the internal benchmark analysis wavers between 0.1 and 0.2, and the number of covariates wavers between 10 to 50, I need between 35 and 70 stops/arrests to achieve desirable statistical power. Not surprisingly, the average of these two values yields an estimated sample size of 53, which roughly aligns with Ridgeway and MacDonald's (2009) original recommendation. For consistency, I use the same 50 stop/arrest cutoff given the supporting results of the power analysis.

Figure 1. Estimated Sample Size for Internal Benchmark Analysis



After sample restrictions, this resulted in an analyzable sample of 3,959 officers based on stop data and 398 officers based on arrest data. The 3,959 officers based on stop data make up about 63% of all 6,290 officers in the general use dataset on stops before any data cleaning. The 398 officers make up about 7% of all 6,015 officers in the general use arrest dataset as well. Based

on these data, CPD officers conducted, on average, 170 stops, with some officers having conducted as many as 1,574 stops over the four years. CPD officers conducted an average of 74 arrests during these four years, with some having conducted as many as 244. Importantly, these estimates do not represent the stop and arrest totals for all officers in the CPD; they are based on the analyzable data after excluding low-discretion stops and arrests, officers listed as anything other than "Police Officer" on their shift assignment, and officers that did not conduct more than 50 stops/arrests. More details on the comparability between these samples and the general officer population of CPD will be discussed in the next chapter.

Beat Boundaries and Racial Composition

Like other large metropolitan police agencies, CPD is ecologically designed around the contours of the city and variations in public need. As shown in Figure 2, police districts represent the largest set of geographical boundaries, which are composed of several sectors, and then a set of smaller police beats. As depicted more closely in District 9 of Figure 3, CPD districts have several police beats, each of which varies in shape and size. While Chicago is a racially diverse city, Figure 4 shows that there are varying concentrations of racial/ethnic minority populations within and between police beats and districts at a block-group level.

Figure 2. Map of Chicago Police Districts and Police Beats (2012-2015)

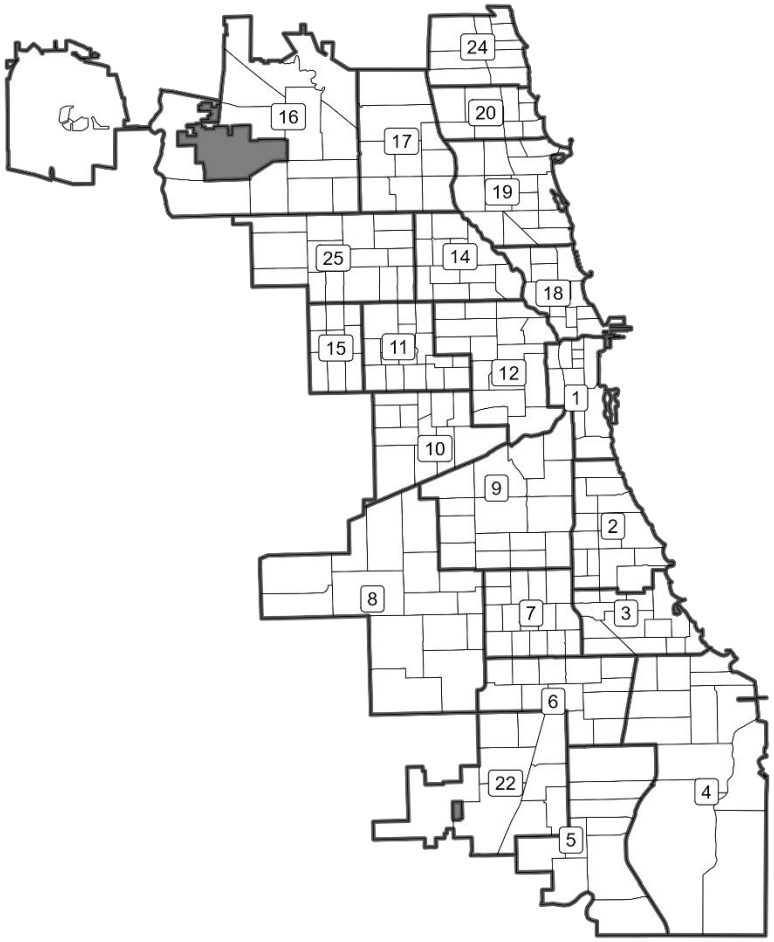
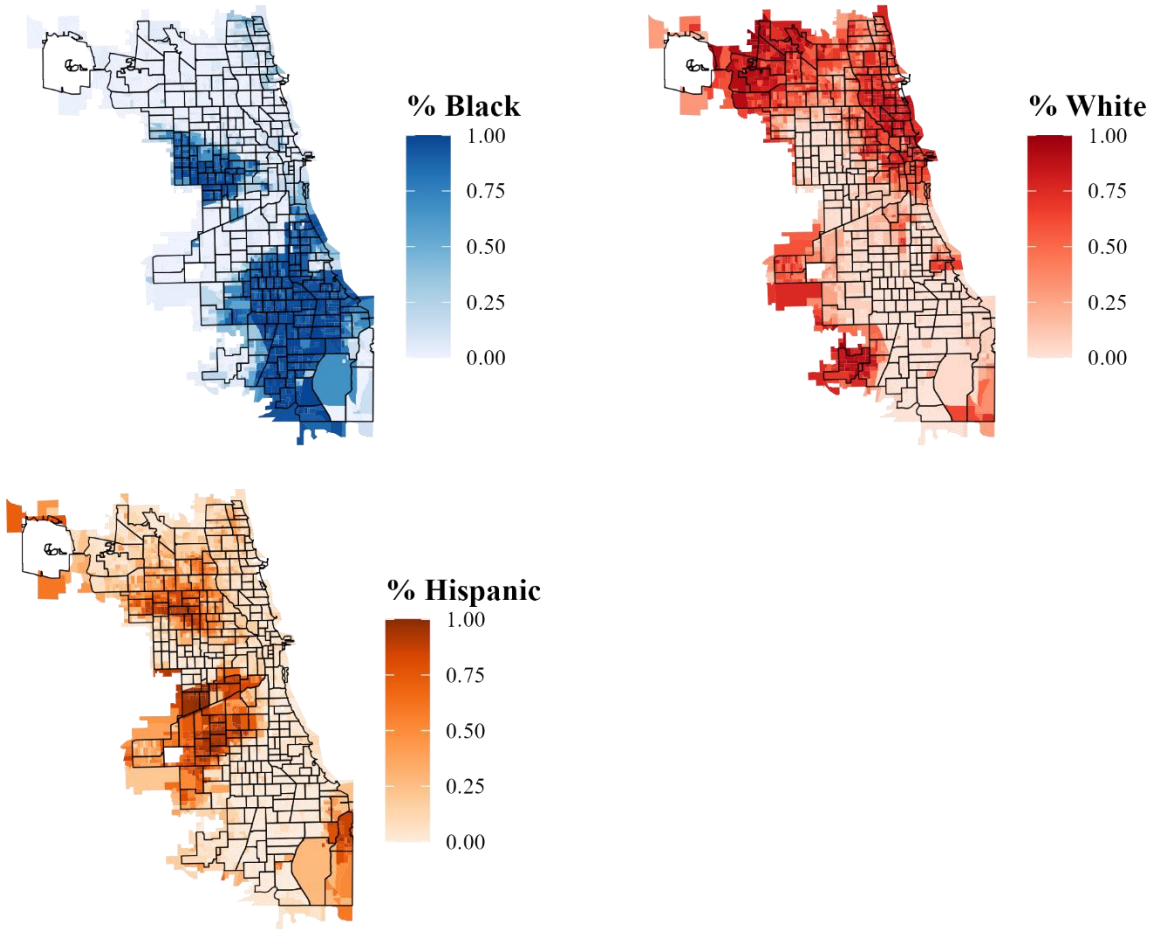


Figure 3. Map of Chicago Police District 9 Police Beats (2012-2015)

Figure 4. Racial/Ethnic Distribution of Chicago Residents



Data source: American Community Survey 5-year estimates (2015)

Accordingly, while it is important to account for differences in what beat assignments officers receive when comparing their behavior, no racial disparity analysis is truly fair in Chicago without also accounting for variation in racial composition at a more micro level. Fortunately, most data on officer stops and arrests have associated geographic coordinates for each incident, which means that incidents can be linked to what block group they occurred in. Those without such information were dropped from the analyzable datasets.[24] This required dropping an additional 36% of arresting officers (N = 151) who did not have at least 50 arrest incidents with geographic

[24] Approximately 81% of arrests and 85% of stops have coordinate data.

coordinates, which resulted in a final analyzable sample of 261 arresting officers. Meanwhile, an additional 8% of stopping officers (N = 319) dropped due to not having at least 50 stop incidents with geographic coordinates, resulting in a final analyzable sample of 3,640 stopping officers.

Accordingly, in this study, racial variation in the risk set of those most likely to come into contact with the police is measured at the block group level. Incorporating information on officers' beat assignments and the block group where they stop or arrest civilians will provide fairer and more accurate officer-level racial disparity analyses. All block group level data measured in this study come from the U.S. Census American Community Survey 5-year estimates (2015). How these data are used in the disparity analyses is outlined in the analytic strategy section to follow.

The Internal Benchmark Analysis

The analysis begins by estimating racial disparities in police behavior among individual officers in the CPD using the internal benchmarking approach originally proposed by Walker (2001) and advanced by Ridgeway and MacDonald (2009). Accordingly, this study follows Ridgeway and MacDonald (2009) by combining three statistical methodologies to construct valid internal benchmarks and thus determine which officers—if any—are potential bad apples. Each step is outlined in Table 6.

Table 6. Steps to the Internal Benchmark Analysis Steps

Step	Description	Reason	Result
1	Matching	Induces an apples-to-apples comparison setting.	Each officer has a matched benchmark of stops/arrests based on stops/arrests engaged by their peers in similar times, places, and contexts.
2	Doubly Robust Estimation	Estimates bad apples while removing residual bias and variance in the matched benchmarks.	Provides preliminary dataset of at-risk officers for racial disparity.

Table 6. (cont'd)

3	False Discovery Rates	Shrinks the risk of False Positives (Type-I Error).	Provides final dataset of at-risk officers.

Creating the Apples-to-Apples Setting: Propensity Score Matching

The first step in the analysis involves creating an internal benchmark dataset for each officer in the agency that consists of stops/arrests made by other officers in similar times, places, and contexts.[25] Accordingly, the internal benchmark works by effectively comparing stops made by an officer in question with stops made by other officers. Internal benchmarks are customized for each police officer by matching the joint distribution of characteristics of stops (i.e., location, time of day, month) conducted by an officer in question with the joint distribution of characteristics of stops made by other officers in the agency. Matching stops by the joint distribution of their temporal, environmental, and situational characteristics ensures that the behavior of an officer in question can be fairly compared against other officers while accounting for important variations in where and when they engage in such behavior. Accordingly, the matching procedure reweights stops of the internal benchmark to make their distributions comparable with those of stops conducted by an officer in question:

$$f(\mathbf{x}|t=1) \;=\; w(\mathbf{x})f(\mathbf{x}|t=0) \qquad \text{Equation (2)}$$

where **x** is a vector of stop characteristics (e.g., location, time of day, month), *t* is a binary indicator for a stop involving the officer in question, and $w(\mathbf{x})$ is the weight function, for which I solve so that the characteristic distribution of stops by the officer in question is equal to the distribution of the benchmark stops.

Equalizing the characteristic distributions of stops through a weight function, $w(\mathbf{x})$, creates

[25] For sake of brevity and clarity, I refer to stops as the primary outcome in the analysis. However, arrests are also examined.

a feasible casual inference model, whereby the racial distribution of stops by an officer in question can be compared against their internal benchmark dataset.[26] For example, I can compare the percentage of stops involving Black civilians for an officer in question to their benchmark dataset of stops only once that benchmark has its stops weighted so that they "appear" quite similar to those conducted by the officer in question. The only other remaining difference between these two stop distributions is whether the stops were conducted by the officer in question, which is reflected by the treatment indicator to create an empirically robust estimate of racial disparity.

In following Ridgeway and MacDonald (2009), the weight function can be estimated by first estimating the propensity score for treatment. The propensity score is estimated using a regression model for binary dependent variables, such as logit and probit models (Long, 1997). In this study, a logistic regression model estimated the propensity of being treated given the characteristics of a stop:

$$\Pr(t = 1|\mathbf{x}) = \frac{\exp(\mathbf{x}\beta)}{1 + \exp(\mathbf{x}\beta)} \qquad \text{Equation (3)}$$

When selecting covariates **x** to predict treatment assignment, Ridgeway and MacDonald (2009) suggest including stop features such as month of year, time of day, day of week, beat assignment, and other characteristics to capture anything that might determine if a stop was conducted by the officer in question (i.e., treatment assignment) that is also jointly related to the stop involving a non-White citizen.

In this study, I included temporal, ecological, administrative, and situational variables in the matching models. Temporal variables include the month, day, and hour in which a stop/arrest took place. These variables were measured continuously because the matching models use a

[26] This hints at the *conditional independence assumption* in matching methods, which assumes that treatment assignment (whether a stop is conducted by the officer in question or their peers) is as good as random given a set of observed variables (Rosenbaum & Rubin, 1983).

nonparametric approach to determine the functional form of their relationship with treatment assignment. In other words, measuring these variables continuously allows the estimation procedure to flexibly match the distribution of each variable between the officer in question and their benchmark dataset. This should generate a smoother covariate balance between the officer in question's stops and their internal benchmark.

Ecological variables include what district an officer was assigned to, where the stop took place and the level of concentrated disadvantage in the surrounding area. Specifically, stops were compared based on their exact geographic coordinates from which they occurred. Measures for which block group and the level of concentrated disadvantage of that block group are included as covariates as well given that I can track where each stop/arrest occurred exactly in Chicago. Here, concentrated disadvantage was measured as a factor score using principal-axis factoring based on the following items: percent of the residential population below the poverty line, percent on public assistance, percent unemployed, percent of female-headed families, and percent Black or African American (Sampson et al., 1997).

To account for potential variations in what duties officers were performing on patrol, two administrative variables were included in the matching models. The first, police beat type, is measured dichotomously as whether an officer was assigned to work in a standard or non-standard beat location. As mentioned above, non-standard locations may involve specialized patrol assignments or duties that otherwise would not be identified in standard beat locations. The second variable, operations calendar code, identifies what squad rotation an officer was working on a specific assignment. These assignment codes are attached to 39% of all assignments, with each code denoted by a phonetic alphabet letter (A, B, C, D…). Including this categorical measure in the models will help account for potential variations in assignments that are attributable to what

rotation an officer works in CPD's operational calendar.

Lastly, I included a measure for stop reason that officers provided when conducting a stop and the type of arrest an officer made. The stop reasons take on five broad categories in part due to variations in agency-based codes (traffic stop, suspicious person, investigatory stop, gang-related, other), which will help account for the confounding influence of why officers made stops in the first place when conducting the internal benchmark analysis. The arrest reasons were coded into 5 categories as well (violent, drug, property, traffic, other). Table 7 provides a complete list of the stop characteristics, **x**, incorporated into the propensity score model.

Table 7. Features for Propensity Score Matching Model

Feature	**Measurement**
Temporal	
Month of Year	Continuous
Day of Week	Continuous
Time of Day	Continuous
Ecological	
Police District	Categorical
Block Group Disadvantage	Continuous
Latitude and Longitude	Continuous
Administrative	
Police Beat Type	Dichotomous
Operations Calendar Code	Categorical
Situational	
Stop Reason	Categorical
Arrest Type	Categorical

Researchers can also include interactions between these covariates when theoretically relevant for explaining treatment assignment. As discussed by Ba and colleagues (2021), CPD officers can be coarsely compared based on what beat and shift they were assigned to during a specific day and month of the year. This yields a high-dimensional interaction between the stop features month×day×shift×beat that they refer to as "MDSB." Comparing officers based on similar MDSBs creates a situation where officers are behaving at the same times and places and have

similar populations at risk for their contact.

Comparing officers based on similar beat assignment types ensures that officers performing similar duties and functions are adequately compared against in addition to their geographic beat location. This is ideal for an internal benchmark analysis but I contend that it is not sufficient on its own because it fails to account for other factors that differentiate officers' stop behavior within MDSBs, such as where officers conducted a stop in a beat, what time of day they stopped someone, and what the ecological context of that stop location was when they chose to stop someone. Accordingly, additional interactions between other variables of interest may be required to produce fair comparisons in the benchmark. Fortunately, interactions between covariates in the matching models I estimate are iteratively tested based on their predictive capacity to best fit each officer's stop and arrest data distributions.

Unfortunately, one of the unique challenges to estimating propensity scores through logistic regression is that the model performs poorly when the functional form of its predictor variables is nonlinear, as is the case with high dimensional interaction terms (Huntington-Klein, 2021).[27] As recommended by Ridgeway & MacDonald (2009), an optimal solution to this problem is to use machine learning methods, such as boosted logistic regression models. According to McCaffrey et al (2004, p. 8), "[b]oosting is a general, automated, data-adaptive algorithm that can be used with a large number of pretreatment covariates to fit a nonlinear surface and predict treatment assignment." Unlike traditional logistic regression models, generalized boosting models (GBM) produce probability estimates for treatment assignment without suffering from the prediction errors induced by incorporating high-dimensional interactions and many covariates in

[27] While functional form issues plague variance and bias in the logistic regression model predictions, others note that having many covariates can quickly exhaust the degrees of freedom in the model as well (McCaffrey, Ridgeway, & Morral, 2004)

the model (McCaffrey et al., 2004). Moreover, boosting allows for a data-driven process to determine the functional form of the relationship within different values of a predictor variable and treatment assignment. This means that the matching model can allow for more flexible, nonlinear relationships that may better characterize an officer's stop behavior given the observed variables at our disposal.

Using a boosting approach to matching officer stops offers several additional advantages when compared to the MDSB approach used by Ba and colleagues (2021). It offers a data-driven approach to probing distinct interactions between variables to maximize the accuracy of the matching model for each officer in question. Accordingly, rather than relying on the MDSB, this study takes advantage of the intuition behind boosting to model many complex variable interactions and develop representative comparisons between officers instead of forcing a strict comparison based on a single high-dimensional interaction.

The added benefit is that concerns about bias are further mitigated through the matching procedure based on other important determinants of an officer's stop behavior, which may not be achieved using the MDSB alone. This includes consideration of temporal variation in stops and arrests within shifts, which may reflect variation in patrol patterns and human routine activity patterns that generate differences in officers' stop and arrest behavior of non-White civilians. Failing to account for this micro-temporal variation within shifts could lead to biased or inefficient estimates of racial disparity. The approach utilized in this study also considers more micro-level variation in patrol activity within beat assignments.

Another distinct advantage over the MDSB approach is that the matches created in this procedure are interrogated based on their accuracy in generating a fair benchmark for an officer in question. Comparing officers based on their beat assignments and temporal factors (month, day,

shift) ensures comparisons made are based on officers performing similar duties under a similar set of circumstances. However, unlike the approach in this study, the MDSB approach assumes every comparison made within an MDSB is perfectly generated. Yet, for the reasons mentioned above, MDSB comparisons alone may not be perfect. Accounting for potential imprecision in the comparison is key to conducting a fair and unbiased assessment of racial disparity and is thus a foundational component of the approach used in this study.

Given the many advantages of GBM, the question becomes how does it estimate the propensity score differently than a traditional logistic regression model? The answer is iterative; however, to better understand how this works it is best to adopt the notion that not one model can effectively predict treatment assignment, but an ensemble of models can (Natekin & Knoll, 2013). At the heart of a GBM is a series of iteratively derived regression trees predicting treatment assignment on a given covariate that are added together to estimate a global propensity score. Therefore, GBM is an iterative process that starts with estimating the propensity score using a logistic regression model, checking which observations were poorly predicted, and re-estimating the model iteratively while weighting poorly predicted observations greater to enhance their prediction, which in turn enhances the matching model accuracy over many iterations. The algorithm initially sets the propensity score, $\mathrm{p}(\mathbf{x})$, to a constant equal to the proportion of stops in the dataset from the officer in question, which is measured through $g(\mathbf{x}) = log(\frac{p(\mathbf{x})}{1-p(\mathbf{x})})$. Each model iteration follows with a computed propensity score and its residual, which are then used to identify small adjustments to add to this initial estimate and inform subsequent model iterations and predictions that maximize the log-likelihood function of the GBM.

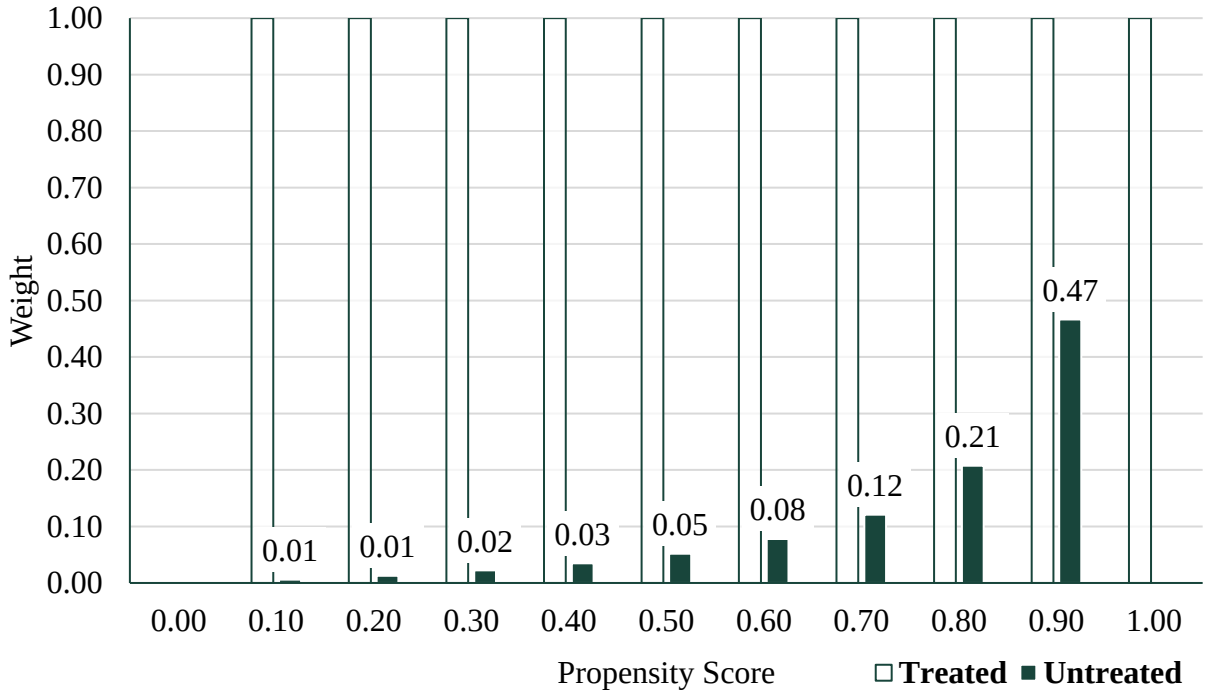
With the propensity scores, p_i, estimated from the GBM procedure, weights are applied to each stop for the internal benchmark analysis. Recall that the goal is to equalize the distribution of

an officer's stops with their internal benchmark per Equation (2), which can be achieved by weighting stops through the formula:

$$w(\mathbf{x}) = \begin{cases} p_i = 1, & t = 1 \\ \dfrac{p_i}{1 - p_i}, & t = 0 \end{cases} \qquad \text{Equation (4)}$$

Weighting internal benchmark stops by their inverse probability of treatment ensures those stops having features that are quite similar to the characteristics of stops by the officer in question will have propensity scores near 1 and therefore receive weights near 1. Weights decay as dissimilarity increases by way of the propensity score for the untreated stops, which is visualized in Figure 5. As described by Hirano et al (2003), inverse probability weights provide one of the most precise ways to estimate casual effects in propensity score matching procedures under conditions with strong statistical power.

Figure 5. Weight Distribution by Treatment Status and Propensity Score



One of the challenges to matching procedures is the need to make features of stops

conducted by each officer in question comparable to features of stops conducted by other officers that form their internal benchmark. As indicated in Figure 5, any benchmark stops with propensity scores equal to zero are dropped from the analysis by the weight calculation in Equation (4). For example, if an officer in question was the only officer assigned to a particular section-8 housing complex, and was not assigned to any other places, they would not have an eligible set of stops to be compared against. In other words, all candidate stops would have a propensity score near or at zero and thus be dropped from the internal benchmark. Therefore, officers whose stops have no valid comparisons must be omitted from the analyzable sample. More broadly, this refers to the issue of *common support* in matching procedures, wherein analysts make sure there are appropriate control observations to match the treated observations on (Caliendo & Kopeinig, 2008).

In alignment with prior research using GBM and internal benchmarking (e.g., Nguyen & Ridgeway, 2023; Ridgeway & MacDonald, 2009, 2014; Ridgeway et al., 2020), I deployed an iterative process to check and satisfy the common support condition. First, I tested the initial balance between stop features for the officer in question with the distribution of stop features generated from their internal benchmark after a first round of matching through GBM. In borrowing from Ridgeway and MacDonald (2009), if an officer in question had at least one stop feature that differed considerably from that of their internal benchmark, then the matching procedure may not have generated a reliable comparison group. Put simply, if an officer has a stop characteristic that differs from their internal benchmark, such as the level of concentrated disadvantage in the block group that they conduct stops, and this is associated with the chances of a Black civilian being stopped by police, then the benchmark will produce a biased estimate of disparity. Accordingly, having any single stop feature differ between an officer in question and their benchmark will diminish the validity of the benchmark altogether, thus I need to identify if

and when this happens to ensure unbiased estimates of racial disparity.

I use a measure of standardized percent bias to capture the imbalance between the marginal distributions of stop characteristics for an officer in question and their benchmark. This estimates the distance in marginal distributions of the stop features between treated and untreated observations based on the square root of the average of sample variances in both groups. Standardized percent bias greater than 10% on any individual feature may be of great concern and thus a sign of poor match quality (Caliendo & Kopeinig, 2008). Based on this information, I then flagged any officer whose benchmark poorly matched their stop behavior.

In the next step, I explored each officer's stop behavior to determine whether there were any stops in which a reliable benchmark could be generated. Recall that in the original procedure, all stops conducted by an officer in question are analyzed when estimating their benchmark (see Equation 4). However, to induce comparability on at least some of their stops, I restricted their sample of stops to only those in which they are not unique to a given stop feature. In following the earlier example, if an officer conducted stops primarily in a section-8 housing complex that no other officer works in, there is no way to generate a set of comparison stops; thus, I omit all those stops while retaining any and other stops made by that officer that occur in places where other officers have made stops as well. This provides at least some evidence of common support but does not guarantee an improvement in comparability.

Accordingly, I then re-estimated each of these officers' benchmarks and reassessed their measure of standardized percent bias to determine if it fell within the acceptable threshold. If not, I repeated this process until the officers' benchmark yielded an acceptable measure of percent bias. Officers whose number of stops never met the threshold for percent bias or did not have enough stops to conduct the internal benchmark were removed from the final sample of officers. This

consequently reduced the total sample of officers in the study and was done to ensure that officers were fairly assessed through accurate benchmarks.

Reducing Bias in Identification: Doubly Robust Estimation

Estimating racial disparities through the proposed internal benchmark analysis involves calculating the racial distribution of stops conducted by the officer in question and the weighted distribution of stops by their internal benchmark. A weighted bivariate logistic regression model provides an optimal approach to estimate this difference:

$$\Pr(Nonwhite = 1|\mathbf{x}) = w\frac{\exp(\beta_0 + \beta_1 t)}{1 + \exp(\beta_0 + \beta_1 t)} \quad \text{Equation (5)}$$

where the outcome of interest is whether the stop involved a non-White citizen and *t* is a binary indicator of whether the stop was conducted by the officer in question (*t* = 1). Propensity score weights, *w*, were applied to all stops per Equation (4), and the standard error of β_1, which estimates our treatment effect for racial disparity, was calculated using a sandwich estimator to account for the weights (Ridgeway & MacDonald, 2009). Simply put, β_1 indicates whether the officer in question engaged in more or less stops involving a non-White civilian relative to their peers.

For example, if 72% of a target officer's stops involve Black civilians whereas only 45% of stops involve Black civilians among their internal benchmark, this would likely indicate a racial disparity. A logistic regression model estimates a statistically significant z-statistic for β_1 based on a two-tailed null hypothesis test, which will confirm or reject that the officer in question stops a proportion of Black civilians that is significantly different relative to their peers in similar times, places, and contexts. However, this estimate will be biased to the extent that the propensity score weights fail to equalize the stop characteristic distributions between the target officer and their internal benchmark. This could be due to a lack of balance on stop characteristics, omitted variable bias in Equation (5), or both. Analysts can leverage a doubly robust estimation procedure to

assuage these shortcomings to the estimation procedure.

Doubly robust estimation is a way to adjust the regression model used to estimate the treatment effect in Equation (5) by incorporating covariates that might explain variation in the probability of a stop involving a racial minority group. As recommended by Ridgeway and MacDonald (2009), all stop covariates that were used to compile the characteristic distribution (i.e., month, day, beat) were incorporated into the doubly robust estimation. The internal benchmark analysis can provide robust results so long as either the original propensity score model or the doubly robust estimation is properly specified (Huntington-Klein, 2021).[28]

Minimizing Type-I Error: False Discovery Rates

Conducting the first two steps of the internal benchmark analysis yields a z-statistic for each officer in the agency indicating whether they had a significantly different proportion of stops involving racial minorities relative to their peers who conducted stops in similar times, places, and contexts. However, the conventional method for testing their statistical significance is inappropriate given that I am testing hundreds or perhaps thousands of values of z simultaneously instead of just one. To understand the scope and nature of the problem, it is beneficial to recall how a standard null hypothesis test is used to estimate the statistical significance of a z-statistic.

Recall that analysts reject the null hypothesis that an officer is *not a bad apple* (H_0: $\beta_1 = 0$) when the probability of observing the z-statistic obtained given the null is true is less than or equal to 5% of the time—for a standard two-tailed test. Analysts arrive at this conclusion by first establishing an acceptable significance level, which is dictated by a maximum acceptable probability of rejecting the null hypothesis when it is true (i.e., Type-I error) and is traditionally set at $\alpha = .05$. Using this information, analysts specify a rejection rule that determines when the

[28] See Waernbaum (2012) for an applied overview of doubly robust estimation within the context of matching models.

null hypothesis should be rejected, which involves determining whether the absolute value of a z-statistic is larger than a critical value. The critical value is determined by both the Type-I error rate and the distribution of z-statistics. The critical value is chosen so that the probability of a Type-I error is 5% ($\alpha = .05$), which is calculated as the $100(1- \alpha/2) = 97.5^{th}$ percentile on a standard normal distribution N(0,1).[29] All absolute values of z in the region of the normal distribution that fall outside this percentile are deemed statistically significant. Calculating this difference as a value gives us the *p*-value, which is 1-Φ(observed z-statistic). If the *p*-value exceeds the significance level ($p > .05$), analysts cannot reject the null hypothesis given it is true less than or equal to 5% of the time.

Analysts can use the standard null hypothesis test to determine whether an individual officer deviates significantly from their benchmark and is, thus, a *bad apple*. However, recall that each test has its own Type-I error probability, which will be compounded when many officers are tested simultaneously. To see how this works, say I have 50 officers in the study setting, which means I have ($n = 50$) z-statistics to test simultaneously. This is because each officer will have their internal benchmark and thus their own estimate of deviation from that benchmark, β_1, derived from Equation (5). Assume Type-I error for these tests is set at 5% ($\alpha = .05$). The chances of identifying at least one officer as a bad apple by random chance yields the formula:

$$\begin{aligned} P(at\ least\ one\ bad\ apple) &= 1 - P(no\ bad\ apples)^{\mathrm{n}} \\ &= 1 - (1 - 0.05)^{50} \\ &\approx 0.92 \end{aligned} \qquad \text{Equation (6)}$$

where the chance that at least one officer is falsely identified as a bad apple is 92% if I test 50 z-statistics at once. The chance of obtaining a false positive only increases as the number of officers

[29] We use the standard normal distribution for the null hypothesis because it reflects the *expected* distribution of z-statistics had they been sampled randomly up to an infinite number of times (Wooldridge, 2015).

being tested simultaneously increases (Shaffer, 1995).

Although the exact costs of falsely claiming an officer engaged in racially-disparate police behavior are unknown, soft organizational injustices and labeling could have downstream consequences for officers' self-legitimacy, productivity, and job satisfaction (Ashforth & Humphrey, 1997; Wolfe & Nix, 2017). Accordingly, an alternative approach is necessary to estimate the significance of these z-statistics that can control for the potential compounding risk of Type-I error that arises with large-scale hypothesis testing.

One large-scale hypothesis testing approach is to identify bad apples by estimating the probability that an officer exceeds their benchmark given the z-statistic they were assigned. If an officer has a high probability of exceeding their benchmark given the z-statistic they were assigned, then it stands to reason that they are a bad apple. Ridgeway and MacDonald (2009) estimate this probability using the formula:

$$P(bad\ apple|z) = 1 - P(not\ bad\ apple|z) \qquad \text{Equation (7)}$$

where the probability that an officer exceeds their benchmark given their z-statistic, $P(bad\ apple|z)$, is solved by identifying the local false discovery rate (FDR), or the probability that they do not exceed their benchmark given their z-statistic, $P(not\ bad\ apple|z)$.[30] Assuming a solution exists for the local FDR, analysts can determine what the probability is that an officer is a bad apple given their z-statistic. According to Ridgeway and MacDonald (2010), any $P(bad\ apple|z) \geq .50$ would indicate the presence of a bad apple, or what could be roughly described as an officer's significant deviation from their internal benchmark.[31]

[30] This is not to be confused with the global false discovery rate, which is defined by Benjamini & Hochberg (1995) as the proportion of falsely rejected tests among the sample of tests in which the null was reject by the analyst.
[31] Interestingly, this threshold equalizes the chances of Type-I and Type-II error by implying that the cost of failing to identify a bad apple equals the cost of flagging a not-bad apple. The implications and potential adjustments to this threshold assumption are addressed in a later section.

In following Efron's (2004) work on large-scale hypothesis testing, the FDR is solved through the equation:

$$P(not\ bad\ apple|z) = \frac{f(z|not\ bad\ apple)f(not\ bad\ apple)}{f(z)} = \frac{f_0(z)f(not\ bad\ apple)}{f(z)} \quad \text{Equation (8)}$$

where $f(z)$ is the observed density distribution of z-statistics for all officers, and $f_0(z)$ is the distribution of z-statistics of all officers that are not bad apples. Fortunately, $f(z)$ is the density distribution of the z-statistics obtained from the propensity score matching and doubly robust estimation procedures. An exact estimation of $f(z)$ is found by fitting a natural spline to the histogram counts of the observed z-statistics.

As for $f_0(z)$, it represents the empirically derived null distribution N(δ_0, σ_0) that is based on the density distribution of the observed z-statistics, $f(z)$. Traditionally, this is set at standard normal N(0, 1) in single-case null hypothesis testing (see footnote 30). In this large-scale hypothesis testing situation, I need to reduce the risk of Type-I error by re-estimating the null distribution because the standard normal distribution tends to be narrower (less conservative) than what is generally appropriate for multiple hypothesis testing (Goeman & Solari, 2014). Efron (2004) proposed estimating the empirically derived null distribution, N(δ_0, σ_0), by fitting the curve of the density distribution of $f(z)$ to the histogram counts by Poisson regression and obtain the center δ_0, and half-width of the central peak σ_0.

Next, I need to re-estimate the power in Equation (8), which is set by $f(not\ bad\ apple)$. Normally this is set at 80% in a single-hypothesis test, but doing so will increase the chances of Type-I error in large-scale hypothesis tests (Efron, 2004). Therefore, power must be specified to more appropriately reflect the distribution of officers who are potentially not bad apples in the

agency. In other words, I need to determine what is the acceptable probability of an officer not being a bad apple. Standard practice is to set the power at 90%, with the justification that "large-scale hypothesis testing is focused on identifying a small percentage of interesting cases that deserve further investigation" (Efron, 2004, p. 97). Fortunately, this operates in the spirit of the internal benchmark analysis and the bad apple conceptual framework, which assumes that a few officers contribute to most of the problematic behavior in a police agency (Chalfin & Kaplan, 2021; Sherman, 1978; Walker et al., 2001).

With all three parameters obtained for Equation (8), I can estimate the local FDR to determine which officers have a high probability of exceeding their benchmark given their z-statistic. Any officer with a local FDR greater than .50 is deemed to be not a bad apple whereas all those with a local FDR below .50 would be flagged as a bad apple. One tangible benefit to the local FDR approach is that it allows for a null hypothesis test that is based on the observed distribution of z-statistics that effectively equates to the actual population. The extent to which this distribution differs from that of the expected standard normal distribution will contribute to inaccurate significance tests in the benchmark analysis.

Although extensive empirical support exists for local FDR procedures in large-scale hypothesis testing, there is one testable assumption that shapes the potential results regardless of its empirical rigor. Analysts must assume what the acceptable probability of an officer being a bad apple is given their z-statistic, or inversely what is the highest FDR they are comfortable with (Ridgeway & MacDonald, 2009). More broadly, this highlights the tradeoff between Type-I error vs Type-II error—with the current cutoff set to have them equal (see footnote 32). This threshold has received little theoretical attention given its novelty to research on racial inequalities in policing. In such cases, it is more appropriate to report the results based on the bounds of these

thresholds (Christensen & Connault, 2023).[32] That is, when there is uncertainty about the assumptions made in the FDR procedure, it is better to analyze all probable values derived from these assumptions.

The results are reported from the local FDR procedure assuming the FDR must be at or below .50, .25, and .05 for an officer to be flagged as a bad apple. Shrinking the FDR cutoff increases the chances of Type-II error, thereby weighing the costs of falsely identifying a bad apple more so than not flagging one. In this case, the expected number of bad apples will decrease as the FDR cutoff shrinks, thereby yielding a smaller but more definitive sample of bad apples. Collectively this provides a bounds approach design that the results will be reported on.

The Policy Exploration

Given the practical challenges of testing and remediating racial disparities in policing, the following policy-relevant questions are designed to shed some insights into ways that police agencies could attempt to address their disparity problem. Importantly, these questions can only provide indirect evidence in support of (or against) the policies they speak to. This is because none of the explored policies are tested in real settings, rather I use data to test questions related to these policies using previously observed behaviors by CPD officers. Furthermore, when interpreting the results from these analyses, mitigating racial disparity is referred to as the agency-level goal that the analysis and policy address. However, the immediate outcomes in the analyses do not directly measure agency-level racial disparities because their associated tests lack validity and can be misleading. It is for these reasons that the outcomes are interpreted within the context of an agency's goal to mitigate racial disparities by suppressing or quelling the behavior of those officers

[32] Similar applications exist in time series analysis, where analysts can face considerable uncertainty about the properties of an observed time trend. In these cases, their causal inference models will suffer from reliability concerns. Using a bounds approach allows analysts to test their model across all potential values that which they are uncertain about to better understand their findings (Webb et al., 2019).

who were previously identified as engaging in racially-disparate police behavior.

Officer Diversification

One of the central contributions of this study is to test police reforms that lack concrete empirical evidence. Recent debates surrounding police reforms and racial discrimination center on resolving the problem from within, primarily through training the current stock of officers and hiring new ones to better meet the challenges that lie ahead. Lacking, however, is strong empirical evidence of the effectiveness of these practices for mitigating racial disparities in a police agency. Accordingly, this analysis explores whether officers with under-represented gender and racial statuses engage in more racially equitable stop and arrest behavior than their White and male colleagues—all while restricting attention to only those officers who were not found to have engaged in racially-disparate police behavior.

The diversification analysis involves a three-part estimation procedure. First, I conduct a series of internal benchmark analyses for each officer flagged as an outlier for racially-disparate behavior in the agency. Specifically, I construct four distinct internal benchmarks for each outlier. Each benchmark differs on one key demographic factor according to a diversification policy of interest, such that one benchmark is comprised of stops/arrests conducted by all Black officers, White officers, female officers, and male officers and compared to each outlier officer in question. It bears noting that the benchmarks themselves are comprised solely of officers who were not found to be outliers in the original analyses. Having a disparity-free reference group is what allows the analysis to show whether race and or gender predicts differences in the extent to which disparity-free officers engage in racially equitable stops/arrests.

When making the benchmarks, I use the same matching procedure described earlier, such that stops of each internal benchmark are reweighted to make their characteristic distributions

equal to those conducted by the sample of bad apples through Equation (2). Recall that the weight function, $w(\mathbf{x})$, equalizes the joint stop characteristic distributions, for which I solve through Equation (4) by initially estimating the propensity score for being a bad apple—or the probability of a stop having features **x** involves a bad apple officer $f(bad\ apple = 1|\mathbf{x})$. All stop/arrest covariates **x** used to predict treatment assignment in the original internal benchmark analysis were also used as covariates in the policy exploration to maintain consistency (see Table 8). This allows us to compare the racial distribution of stops conducted by bad apple officers to that of their peers who, when conducting stops in similar times, places, and contexts, were either all White officers, Black officers, male officers, or female officers.

After creating each benchmark for each outlier officer, the next step involves measuring the extent to which an officer's racial composition of their stops deviates from each of their benchmarks (Black officer benchmark, White officer benchmark, female officer benchmark, male officer benchmark) per Equation (5). While there will be a racial disparity in stops or arrests when comparing an outlier officer to each of their benchmarks, the magnitude of this disparity may differ based on the gender or race of the officers who made stops or arrests in each benchmark dataset. This is measured as the percentage point difference in the racial composition of each outlier officer's stops and arrests relative to their internal benchmark, which will aid with subsequently comparing the average of these differences between benchmarks.

Accordingly, in the final step of the analysis, I compare the percent-point differences between Black-White benchmarks and male-female benchmarks across all outlier officers to determine if, on average, White officers or Black officers differ in the racial equity of their traffic stop behavior—after accounting for the fact that each set of stops are conducted by officers who are not outliers and are matched similarly based on the context under which they made their stops.

This is the crux of the policy exploration.

For example, if the all-Black officer benchmark yielded an average difference of 5 percentage points between the proportion of stops involving Black civilians when compared to the sample of bad apple officers, and the all-White officer benchmark yielded an average difference of 2.5 percentage points, the corresponding difference-in-disparities would be 2.5. If such a difference is, on average, significantly different from zero across all outlier officer benchmarks, this would indicate that Black officers engage in more racially equitable stops than their White colleagues (among those who do not racially profile). This is because both benchmarks deviate in their racial composition with the bad apples, yet Black officers do more so than White officers under the same set of circumstances. I conducted a two-sample t-test with 10,000 bootstrap replications to estimate these differences between benchmark groups, where the distribution of observed disparity estimates for each benchmark are compared between race and gender.

Police Beat Assignment

Given the costs, contention, and or resistance that may be associated with encouraging agencies to hire a more diverse workforce, some may look to resolve their disparity problems from within (McCrary, 2007). One solution is to reassign officers in the agency who were identified as bad apples to areas where they have a lower risk of engaging in such behavior (Goncalves & Mello, 2021). Reassigning the bad apples may offer agencies a simple, low-cost solution for their disparity problem by simply reducing the chances that they may stop or arrest non-White civilians. Importantly, this assumes that officers' racially-disparate behavior will change across contexts, though this has yet to be tested; thus, further empirical scrutiny is required before any formal recommendation and implementation. Accordingly, I built a regression model to predict the probability that an officer who was identified as an outlier in the original benchmark analyses

engaged in racially-disparate stop and arrest behavior on a given day in a given beat. The question at hand is whether disparate behavior is influenced by police beat composition, presumably due to the opportunities they create based on the residential population of a police beat.

Here, the outcome of interest is a racial disparity estimate measured for each outlier officer working on day *t* in beat *k.* The unit of analysis is thus measured at the officer-level, however, given how assignments are structured in the Chicago Police Department, data are measured as Officer×Date×Beat (ODB). Table 8 presents an example dataset demonstrating the structure of the data and the variables informing this analysis.

Table 8. Officer Beat Assignment Analysis Dataset Structure Demonstration

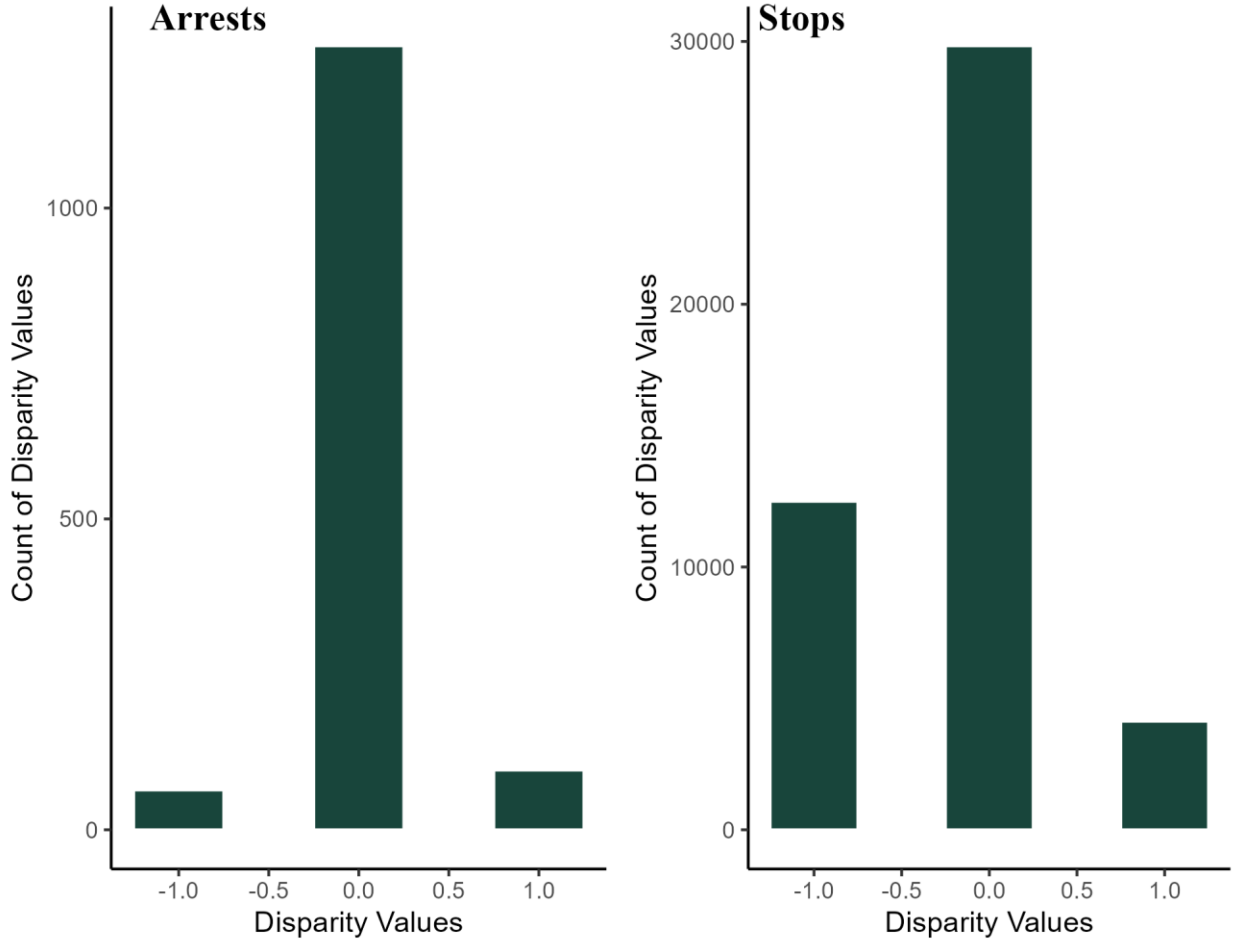
1	2	3	4	5	6	7
Bad Apple Officer	Date	Beat Assigned	% Stops Black by Bad Apple Officer	% Stops Black by Other Officers	Disparity	% of Beat Population Black
Officer A	01/01/2012	0112	30%	30%	0	80%
Officer A	01/02/2012	0112	60%	50%	1	80%
Officer A	01/03/2012	0114	20%	10%	1	30%
Officer A	01/04/2012	0114	10%	20%	-1	30%
Officer B	01/01/2012	2113	80%	60%	1	60%
Officer B	01/02/2012	2113	85%	50%	1	60%
Officer B	01/03/2012	4112	16%	20%	-1	25%
Officer B	01/04/2012	0114	12%	10%	1	30%

To create the disparity estimate, I first observed the proportion of stops/arrests involving Black civilians conducted by a bad apple officer based on the beat and day that they were assigned to work from 2012 to 2015. For reference, this is measured in column 4 of Table 8. I then measured the proportion of stops/arrests involving Black civilians conducted by other officers who were not originally identified as outliers for each beat and day that a bad apple officer also worked (column 5 of Table 8). I finally take the difference between these two proportions to determine whether bad apples arrested/stopped more (or less) Black civilians relative to their peers when assigned to the

same beats on the same days (measured in column 6 of Table 8). Positive values indicate bad apple officers arrested/stopped a greater proportion of Black civilians relative to their peers, and negative values indicate bad apple officers arrested/stopped a smaller proportion of Black civilians relative to their peers.

One interesting feature of the stop and arrest disparity outcomes is their tri-modal distribution, where values below zero tend to be near -1 and values above zero tend to be near +1, and most values tend to be zero (See Figure 6). In both stops and arrests, the disparity indicator for a given police beat on a given day is likely to be 0 with a small proportion of instances in which it is above or below it. This poses a challenge to modeling disparities as they tend to take on one of three values with interpretations for positive and negative values needing to be in reference to 0. Accordingly, to prepare for the regression analysis, the disparity outcome was converted into two separate dichotomous variables (Positive Disparity, Negative Disparity). Here, our positive disparity variable was coded as 0 for all instances where a disparity value was equal to zero (equal representation of Black civilians in bad apple behavior) and coded as 1 for all instances where the disparity value was greater than zero (over-representation of Black civilians in bad apple behavior). A second dichotomous disparity variable, referred to as the negative disparity variable, was coded similarly as 0 for all instances where a disparity value was equal to zero, and coded as 1 for all instances where the disparity value was less than zero (under-representation of Black civilians in bad apple behavior).

Figure 6. Observed Disparity Distribution by Behavior



The key independent variable of interest is the proportion of the residential population that is Black in police beat *k* (see column 7 of Table 8). This is measured by aggregating the American Community Survey 5-year estimates (2015) collected at the block group level.[33] To reduce bias and enhance precision in the regression analysis, I measured the number of violent crimes (measured as Part-1 Uniform Crime Report offenses) at the beat level using publicly available data from Chicago's Open Data Portal. I also included a measure of the day-of-week, month-of-year, and year to account for daily, monthly, and annual variations in disparities. I included a measure of the number of bad apple officers assigned to police beat *i* on day *t*. This will capture any potential

[33] Given the considerable overlap between block groups and police beats, census data were aggregated based on the proportion of the area that which each block group is contained within each police beat.

peer influences related to having more than one bad apple officer in a police beat. Lastly, I included a measure of which outlier officer *i* was being analyzed on date *t* in beat *k* through fixed effects to account for any variation in disparities that may be attributable to the officers themselves. A logit model was estimated using maximum likelihood estimation and is of the general form:

$$y_{ikt} = \beta_0 + \beta_1 \%Black_k + \beta_2 Crime_{kt} + \beta_3 BadApples_{kt} + \beta_4 DayofWeek_w + \beta_5 MonthofYear_m + \beta_6 Year_h + \beta_7 OfficerInQuestion_i + \varepsilon_{it} \quad \text{Equation (9)}$$

where y is a binary variable indicating either a positive or negative disparity for outlier officer *i* in beat *k* on day *t*, with no disparity being the reference point. The primary variable of interest is the proportion of a beat's residential population that is White, which is estimated through β_1. Temporal variation, crime, and the number of bad apples working in beat *k* on day *t* are included as covariates as well. The main policy-relevant question that Equation (9) answers is whether the share of Black residents in a beat is associated with a reduced probability that an outlier officer engages in racially-disparate stop/arrest behavior. As one might expect, just because an officer was found to be an outlier in the original analysis over four years, it does not mean that they are always engaging in racially-disparate behavior every day during every assignment. Accordingly, the focus of this analysis is to see whether on days when they do engage in that behavior, it is more likely to occur in contexts where the residential population is more likely to be Black.

Overview

The overarching purpose of this study is two-fold. The main purpose is to assess whether any officers engaged in a pattern of behavior that would constitute racially-disparate policing based on a leading approach to measuring officer-level disparities. In the next section, the results of this analysis are presented. This begins with an overview of the sample of officers that comprise the

arrest dataset and stop datasets. I then discuss the results of the matching procedure, paying particular attention to those that did not match and reporting the number of officers that were thus omitted from the main analysis. I proceed by discussing the results following the FDR procedure, outlining how many officers (if any) were flagged as having significant estimates of racial disparity after accounting for the elevated risk of false positive detection. I report these results based on based on how much I outweigh the importance of failing to flag a bad apple officer (Type-II error).

I then discuss the demographic makeup of officers flagged as having significant disparity estimates relative to those who did not. This will provide early insights into whether such characteristics were associated with the chances of being flagged as an outlier. Lastly, I explore whether racial disparities are behavior-specific by identifying to what extent officers had significant disparities in their stop and arrest behavior. This will provide unique, data-informed insights into a largely unexplored topic in the etiology of racial discrimination in policing: behavioral invariance in racially-disparate police behavior.

The second purpose of this study is to answer theoretically motivated and policy-relevant questions that seek to address racial inequalities at a micro level. The first set of analyses involves comparing the racial distribution of stops/arrests made by bad apple officers to other officers in the agency that conducted stops/arrests in similar times, places, and contexts but did not engage in disparate behavior and differ on a key demographic feature (i.e., race, sex). I then determine whether officers who did not engage in racially-disparate behavior and have under-represented racial and gender statuses engage in more racially equitable behavior—after accounting for the fact that their stops/arrests are matched similarly based on the context under which they are made.

In the next analysis, I explore whether officers who are at risk for engaging in racially-disparate police behavior are less likely to engage in that behavior when working in jurisdictions

with smaller populations of Black civilians. Building on the situational crime prevention framework, the logic follows that by having these officers work in low-risk settings with fewer opportunities to perpetuate racially-disparate police behavior, agencies can mitigate such behavior. To do this, I constructed a regression model to predict the probability that an officer who was identified as an outlier in the original benchmark analyses engaged in racially-disparate stop and arrest behavior on a given day in a given beat over four years, and whether this probability was lower in police beats with smaller residential concentrations of Black civilians.

CHAPTER 4: RESULTS

Descriptive Statistics of Analyzable Samples

To begin, I highlight differences in the demographic composition of officers that comprise the analyzable and general-use datasets for arrest and stop behaviors. I also compare the demographic composition of officers in these datasets to that of the entire agency based on data from the 2016 Law Enforcement Management and Administration Survey, all of which are presented in Table 9. I do this to underline the extent to which these datasets are representative of all officers in CPD, which is important when contextualizing the findings of the racial disparity analyses.

Table 9. Comparison of Officer Characteristics Between Analyzable Sample and General Sample for Arrest and Stop Data

Arrest Data		Analyzable Data Officer Sample (N = 261)	General Use Data Officer Sample (N = 6,015)	2016 LEMAS Sworn Personnel (N = 11,965)
Race	Black	11%	24%	22%
	Hispanic	28%	24%	23%
	Other	4%	4%	3%
	White	57%	48%	52%
Gender	Female	10%	23%	22%
	Male	90%	77%	78%
Spanish Speaking	No	88%	86%	
	Yes	12%	14%	
Age	Mean	40	42	
	Sd	7	8	
	Min	28	22	
	Max	60	66	
Total Arrests	Mean	71	19	
	Sd	23	20	
	Min	50	1	
	Max	205	243	

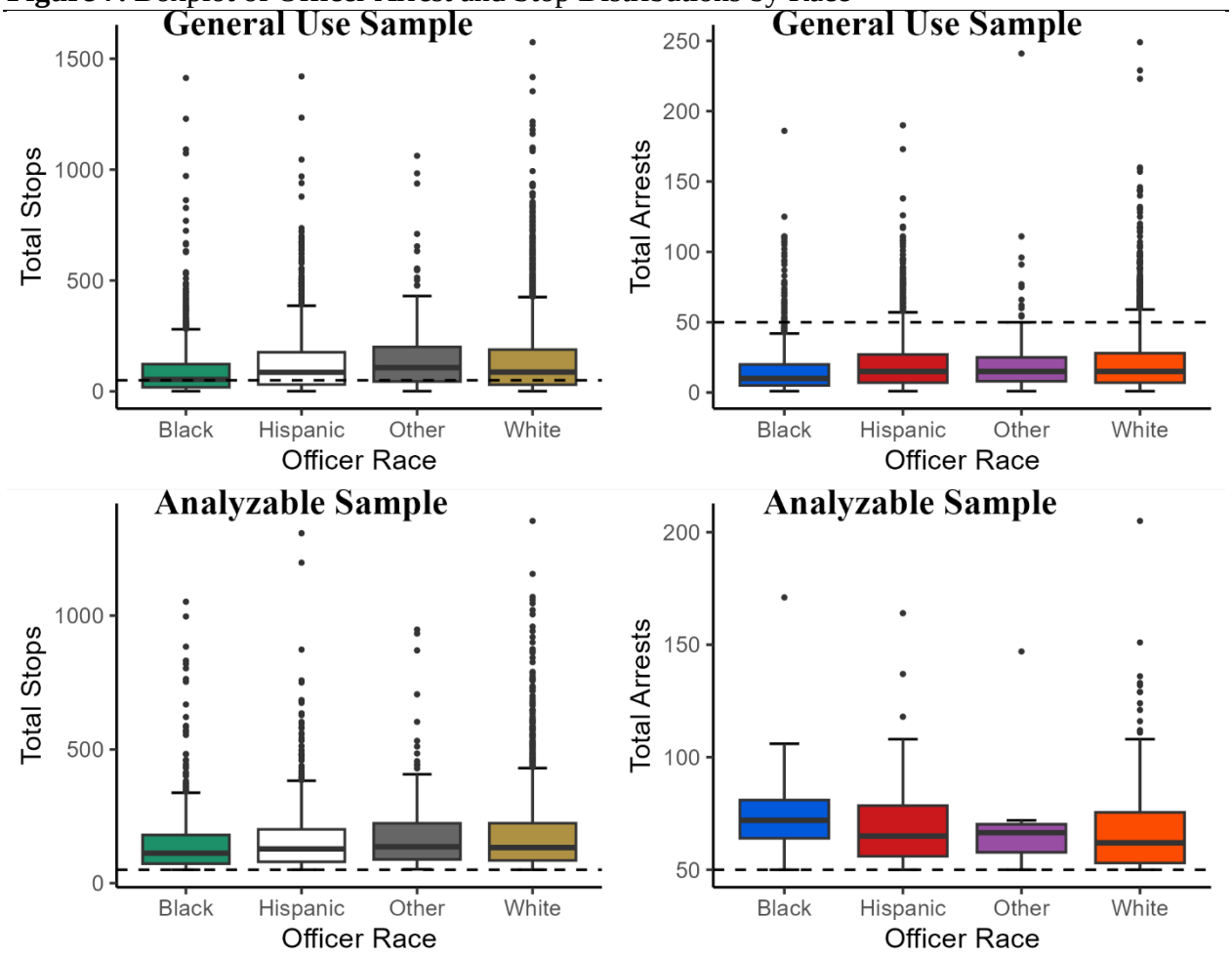
Table 9. (cont'd)

Stop Data		Analyzable Data Officer Sample (N = 3,640)	General Use Data Officer Sample (N = 6,504)	2016 LEMAS Sworn Personnel (N = 11,965)
Race	Black	20%	24%	22%
	Hispanic	25%	24%	23%
	Other	5%	4%	3%
	White	50%	48%	52%
Gender	Female	22%	24%	22%
	Male	78%	76%	78%
Spanish Speaking	No	86%	86%	
	Yes	14%	14%	
Age	Mean	42	42	
	Sd	8	9	
	Min	26	22	
	Max	65	66	
Total Stops	Mean	170	218	
	Sd	140	225	
	Min	50	1	
	Max	1,354	2,028	

Results from Table 9 reveal four main findings. First, the analyzable dataset of arresting and stopping officers is not a perfect representation of the general-use dataset. There were only 19 arrests conducted per officer based on data from the general use dataset, which is far less frequent than the average of 71 arrests per officer in the analyzable dataset of arrests. In contrast, each officer in the general use dataset conducted 218 stops over the study period compared to 170 stops in the analyzable dataset. What this means is that the analyzable sample of arresting officers may be more active than the average CPD officer, while the sample of stopping officers may be less active than the usual CPD officer. More on the implications of these findings will be discussed in the conclusion chapter.

Second, Black officers are generally under-represented among the analyzable sample of arresting and stopping officers when compared to their composition in the general use datasets and among sworn personnel across the entire agency. This should come as no surprise given that Black officers across the general use sample conducted an average of 15 arrests over the study period, which is 25% less than Hispanic officers and White officers ($\bar{x}$ = 20). Meanwhile, Black officers conducted an average of 93 stops, which is 27% fewer stops than Hispanic officers ($\bar{x}$ = 128) and 33% fewer than White officers ($\bar{x}$ = 139). This is presented visually in Figure 7 through a series of boxplots, which display the distribution of total stops and arrests by officer race between the types of datasets and their averages (demarcated by the solid line in each box).

Figure 7. Boxplot of Officer Arrest and Stop Distributions by Race



Note: Dashed line set at 50 arrests/stops

Another finding revealed in Table 9 is that females are under-represented among arresting officers, and slightly under-represented among stopping officers in the analyzable datasets relative to their composition in the general use datasets and among sworn personnel across the entire agency. Again, this is driven by the fact that female officers in CPD conduct, on average, fewer arrests ($\bar{x}$ = 15) and stops ($\bar{x}$ = 111) compared to male officers (arrests $\bar{x}$ = 21, stops $\bar{x}$ = 130).

It is also important to recognize the overall reduction in sample size that is attributed to the sample restrictions employed in this study. As is clear in Table 9 and Figure 7, most officers did not conduct enough arrests to be considered for the internal benchmark analysis, which was disproportionately concentrated among White officers and male officers. In contrast, most officers

conducted enough stops to be considered for the internal benchmark analysis. It is important to recognize these trends in representation as this will contextualize later findings related to predicting who engages in disparate behavior and who engages in more racially equitable behavior.

Matching Procedure Results

When analyzing stops and arrests made by each officer in the sample, the matching procedure was able to create a matched counterfactual dataset of stops and arrests for some but not all officers. Specifically, the matching models identified a suitable benchmark for 53% of arresting officers (N = 139). This is based on all measures of standardized percent bias for each arrest characteristic being less than 10%. This deviation in the KS statistic was primarily due to most of the officers' arrests being made while on beat assignments that few others had, thereby making it difficult to create a comparable set of arrests.

As noted in Table 10, officers with matched benchmarks were a few years younger and made more arrests than their peers who had an unbalanced set of counterfactual arrests when compared to the unmatched officer sample. I investigated the 124 unmatched arresting officers to determine whether omitting their incomparable arrests would yield a more suitable benchmark. This involved omitting arrests conducted by an officer in question whose matched benchmark of arrests had extremely low propensity scores. Namely, all stops conducted by the officer in question whose benchmark arrests had propensity scores among the 1st percentile of all propensity scores in the arrest benchmark were omitted. Benchmark arrests with exceedingly low propensity scores clustered on the arrest characteristics with the least common support; thus, omitting the incomparable arrests by an officer in question will assist in achieving a more suitable benchmark.

Table 10. Comparison of Officer Characteristics between Officers with Matched and Unmatched Arrests and Stops

Arrest Data		Officers With Matches (N = 139)	Officers Without Matches (N = 124)	*t* (X^2)	*p*
Race	Black	9%	14%	(2.73)	0.45
	Hispanic	30%	26%		
	Other	3%	5%		
	White	58%	56%		
Gender	Female	12%	7%	(1.38)	0.30
	Male	88%	92%		
Spanish Speaking	No	89%	87%	(0.28)	0.69
	Yes	11%	13%		
Age	Mean	38	41	-3.53	0.00
Total Arrests	Mean	76	64	4.52	0.00

Stop Data		Officers With Matches (N = 3,010)	Officers Without Matches (N = 630)	*t* (X^2)	*p*
Race	Black	20%	18%	(3.44)	0.33
	Hispanic	25%	25%		
	Other	5%	4%		
	White	50%	53%		
Gender	Female	22%	19%	(2.85)	0.10
	Male	78%	81%		
Spanish Speaking	No	87%	85%	(1.12)	0.31
	Yes	13%	15%		
Age	Mean	41	43	-3.82	0.00
Total Stops	Mean	179	131	8.48	0.00

Notes: All independent sample t-tests are estimated based on 10,000 bootstrapped replications. P values represent bias-corrected estimates based on bootstrapping procedures. All Pearson X^2 tests are estimated using p-values computed by 10,000 Monte Carlo simulations. All two-sample t-test p-values are estimated with 10,000 bootstrap replications.

However, most officers conducted just above the bare minimum number of arrests to be

considered for the initial benchmark analysis ($N \geq 50$) to begin with, which inevitably led to most not having enough comparable arrests to re-estimate their benchmark after omitting many of their incomparable arrests. As such, after accounting for officers with potentially poorly fitting benchmarks, I analyzed 139 officers to determine who engaged in racially-disparate arrest behavior in the subsequent analyses.[34]

The matching procedure for stop data performed much better with 83% of officers (N = 3,010) having an adequately matched benchmark of stops. Much like arrests, officers who did not have an adequately matched benchmark were those who conducted many of their stops while on specific beat assignments that few other officers had. Officers with matched stop benchmarks were a few years younger and had more stops than their peers who had an unbalanced set of counterfactual stops. Again, the only significant difference between the sample of officers with matched and unmatched stop benchmarks was that their mean age was slightly younger, and their stop total was much higher. Accordingly, I analyzed 3,010 officers with suitable stop benchmarks to determine who engaged in racially-disparate stop behavior.

Taken together, results from a series of Pearson X^2 tests and independent sample t-tests shown in Table 10 demonstrate that there were no systematic racial and gender differences between officers who had matched benchmarks and those who did not. This is important because it suggests that race and gender were unrelated to the efficacy of the internal benchmarking procedure—which serves as a strong foundation for assessing statistically unbiased potential differences in race and gender among those who were and were not flagged as an outlier (someone who engaged in disparate behavior) in the forthcoming analyses.

[34] An alternative approach would be to reduce the number of arrests/stops required to be considered for the internal benchmark analysis. However, as noted in Figure 1, a minimum of 50 arrests/stops is required for the analysis to have adequate power. Thus, omitting these officers is required for a robust analysis.

To better understand how the matching procedure worked, Table 11 and Table 12 present results for two different officers that were matched based on the joint distribution of their arrest and stop characteristics, respectively. As noted in each table, the marginal distribution of arrest and stop characteristics made by other officers that comprise their benchmark are weighted per Equation (4) to better approximate the characteristics of stops and arrests made by the officer in question. This is presented in the second column. Unweighted stops and arrests are presented in the third column to show how the marginal distribution changes after propensity score weighting.

The number of observations for each officer in question and their unweighted benchmark indicates the number of arrests or stops they conducted. Accordingly, in this example, the arresting officer made 65 arrests and their unweighted benchmark is comprised of 498 arrests. In the weighted benchmark, the number of observations is measured as its effective sample size (ESS), which captures the increase in sampling variance attributed to weighting the outcome (in this case number of arrests involving Black civilians) based on the arrest characteristics included in the matching model (see Table 7). The decrease in sample size from the unweighted to weighted benchmark shows just how many arrests were reasonably comparable to the officer in question.

Table 11. Percent of Arrests Involving Black Civilians for an Officer-in-Question and Internal Benchmark Sample from the Propensity Score Weighting

			Internal Benchmark	
Arrest Characteristic		Officer in Question (N = 65)	Weighted (ESS = 498)	Unweighted (N = 1,674)
Month	January	9%	10%	9%
	February	8%	8%	9%
	March	9%	10%	10%
	April	12%	12%	9%
	May	9%	9%	9%
	June	12%	10%	9%
	July	6%	7%	8%
	August	15%	14%	7%
	September	3%	5%	8%
	October	9%	9%	8%
	November	5%	5%	7%

Table 11. (cont'd)

	December	2%	3%	6%
Day of Week	Sunday	12%	11%	15%
	Monday	17%	17%	15%
	Tuesday	5%	8%	13%
	Wednesday	18%	15%	13%
	Thursday	20%	20%	15%
	Friday	11%	12%	14%
	Saturday	17%	16%	15%
Time of Day	3-5 p.m.	40%	46%	48%
	6-8 p.m.	48%	42%	38%
	9-11 p.m.	12%	12%	14%
Police District	10	2%	4%	16%
	11	98%	96%	84%
Beat Type	Standard	86%	84%	70%
	Location			
Operational Calendar Code	A	5%	5%	13%
	D	2%	2%	5%
	No Code	94%	93%	82%
Concentrated Disadvantage		1.51	1.53	1.46

Notes: N is the number of arrests conducted, ESS represents the effective sample size, concentrated disadvantage was measured using principal axis factoring according to 5-year annual estimates from the American Community Survey measured at the Block Group level. Time of day is rounded to the nearest 3-hour period.

For each officer in question, the benchmark has almost the same distribution of arrest and stop characteristics, whereas the unweighted benchmarks exhibited considerable differences compared to the officer in question. This shows how the matching procedure can successfully generate a benchmark of stops and arrests that were conducted in the same times, places, and contexts as the officer in question.

Table 12. Percent of Stops Involving Black Civilians for an Officer-in-Question and Internal Benchmark Sample from the Propensity Score Weighting

Stop Characteristic		Officer in Question (N = 730)	Internal Benchmark	
			Weighted (ESS = 716)	Unweighted (N =116,181)
Month	January	11%	13%	9%
	February	13%	13%	9%
	March	8%	10%	11%
	April	8%	6%	9%
	May	10%	9%	8%
	June	14%	12%	8%
	July	8%	6%	8%
	August	7%	8%	8%
	September	3%	4%	8%
	October	6%	6%	8%
	November	8%	9%	8%
	December	4%	3%	6%
Day of Week	Sunday	14%	13%	13%
	Monday	19%	17%	14%
	Tuesday	13%	13%	15%
	Wednesday	14%	14%	15%
	Thursday	18%	20%	14%
	Friday	12%	13%	14%
	Saturday	10%	10%	14%
Time of Day	12-2 a.m.	1%	1%	18%
	12-2 p.m.	1%	1%	11%
	3-5 p.m.	78%	77%	25%
	6-8 p.m.	18%	18%	25%
	9-11 p.m.	3%	3%	22%
District	4	0%	1%	21%
	6	9%	11%	19%
	7	91%	88%	39%
	9	0%	1%	22%
Beat Type	Standard Location	14%	17%	66%
Operational Calendar Code	A	72%	71%	4%
	B	5%	5%	4%
	C	2%	2%	2%
	D	0%	0%	2%
	R	1%	2%	23%

Table 12. (cont'd)

	Other	1%	1%	4%
	No Code	18%	20%	62%
Stop Type	Investigatory Stop	16%	14%	13%
	Other	2%	3%	28%
	Suspicious Person	66%	67%	20%
	Traffic Related	17%	16%	38%
Concentrated Disadvantage		1.86	1.85	1.27

Notes: N is the number of arrests conducted, ESS represents the effective sample size, concentrated disadvantage was measured using principal axis factoring according to 5-year annual estimates from the American Community Survey measured at the Block Group level. Time of day is rounded to the nearest 3-hour period.

The matching procedure also accounts for the geographic distribution of each officer's stops and arrests by matching the exact locations where each officer makes their stops and arrests. Figure 8 and Figure 9 show how stops and arrests conducted by officers in the benchmark receive more weight when they are conducted in closer proximity to those of the officer in question, which is indicated by the size of each point on the map. Here, all arrests (red) and stops (orange) made by the officer in question receive the largest weight per Equation (4). Again, this is to show that the matching procedure not only accounts for officer-specific geographic variation in their behavior between beat assignments but also within them to create highly customized counterfactuals for the internal benchmark analysis.

Figure 8. Distribution of Arrests by Officer in Question and Internal Benchmark

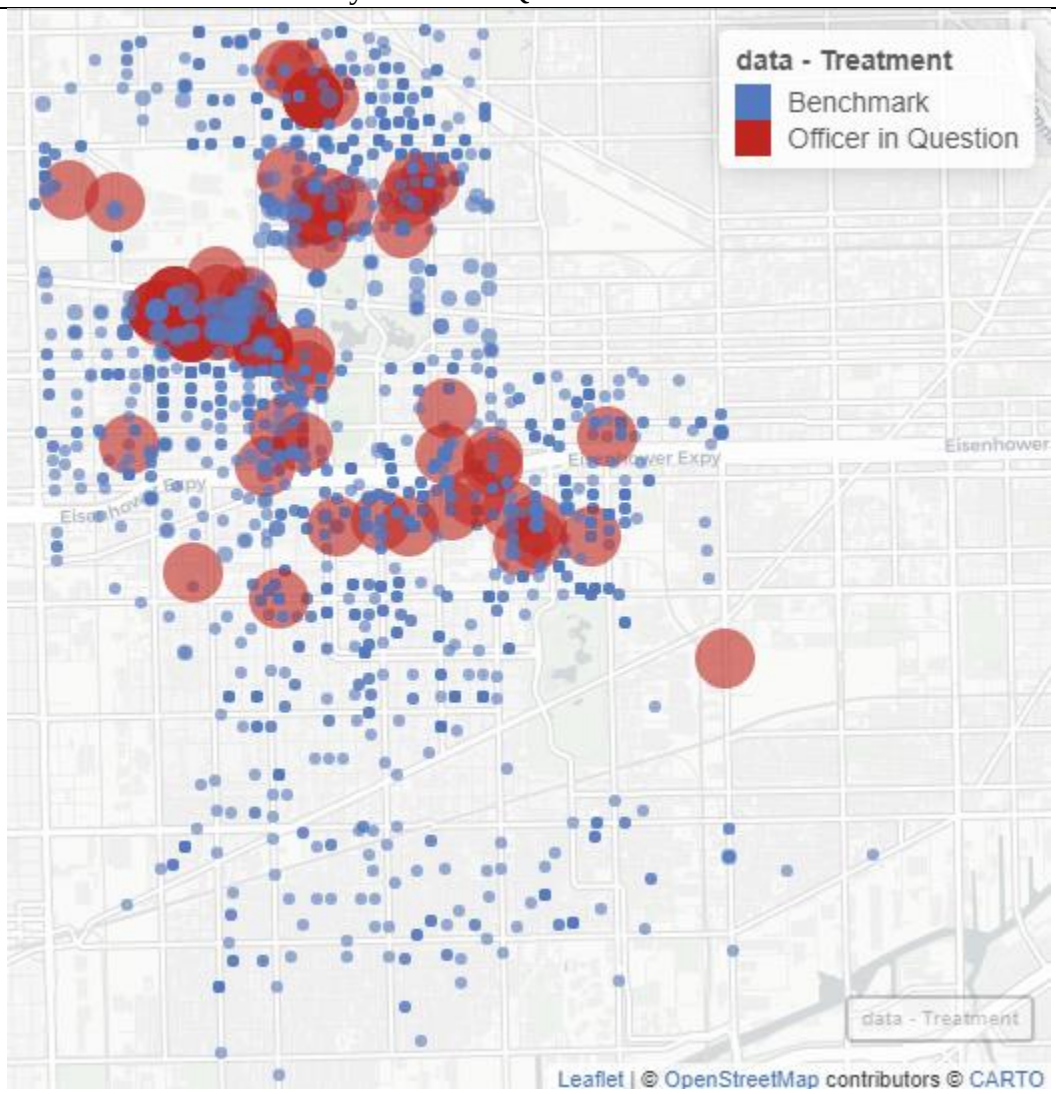
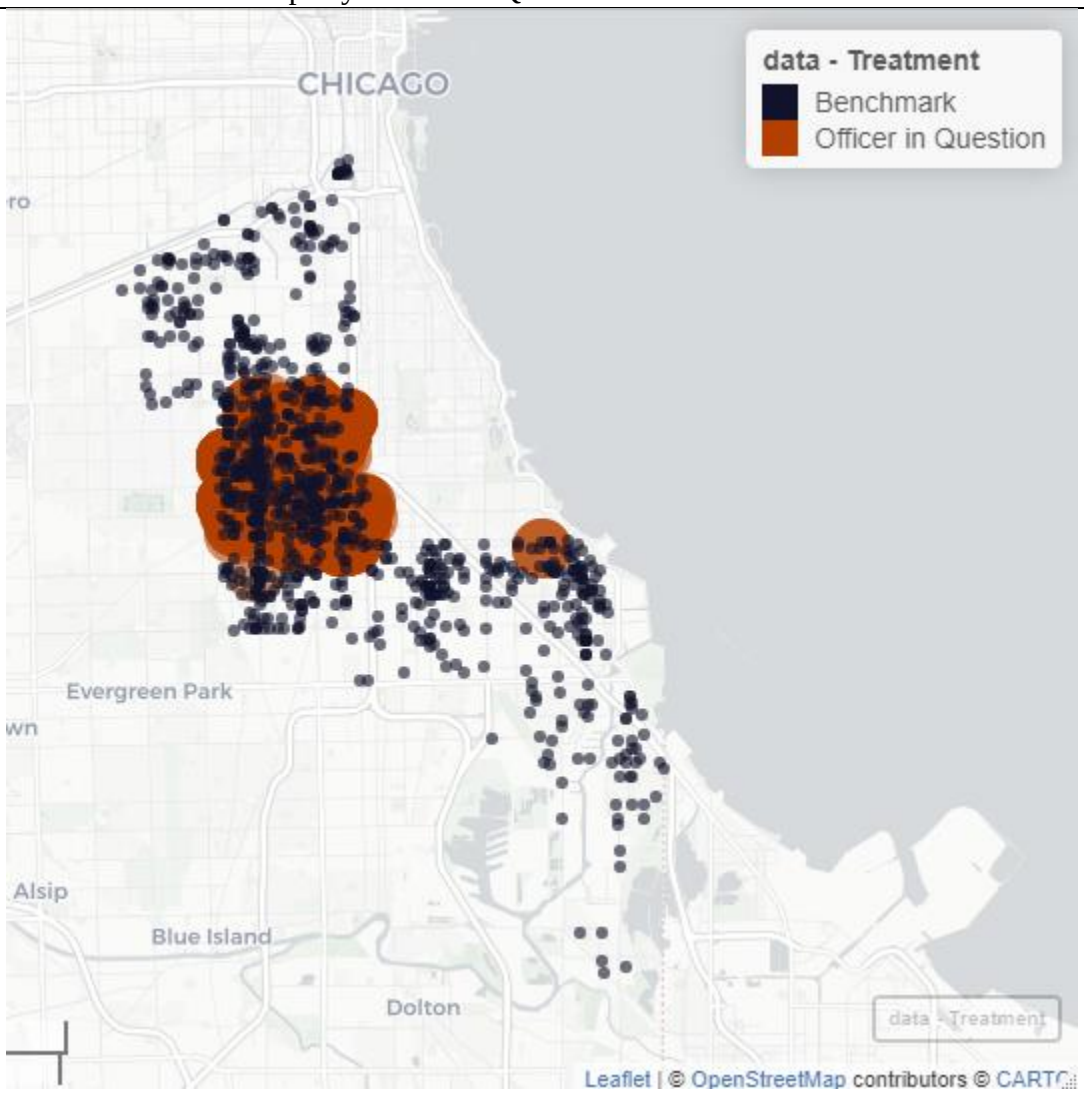


Figure 9. Distribution of Stops by Officer in Question and Internal Benchmark



In summary, these benchmarks provide a set of counterfactual stops and arrests that took place in the same circumstances for each officer in question. Accordingly, I can estimate the extent to which each officer in question has a different proportion of arrests or stops involving Black civilians relative to their benchmark with a high degree of internal validity. For demonstration, these results are presented in Table 13, along with two additional arresting officers and stopping officers for reference. The table reports the proportion of stops and arrests involving Black civilians, with corresponding z-statistics and *p*-values indicating the statistical significance of their differences.

Table 13. Demonstration of Doubly Robust Benchmark Estimation

	Officer		Benchmark				
	% Black	N	% Black	ESS	PPD	*z*	*p*
Arresting Officer 1	98%	65	88%	568.2	10.5	2.15	0.03
Arresting Officer 2	95%	65	92%	105.6	3.0	5.66	0.00
Arresting Officer 3	98%	57	98%	251.3	0.0	0.00	1.00
Stopping Officer A	69%	349	45%	628.6	24.1	7.20	0.00
Stopping Officer B	99%	730	98%	763.6	1.4	2.03	0.04
Stopping Officer C	84%	75	84%	3,477.2	0.2	0.04	0.97

Notes: N is the number of arrests or stops conducted by the officer in question, ESS represents the effective sample size of each officer's benchmark. (PPD) refers to the percentage point difference between the proportion of stops/arrests involving Black civilians and their internal benchmark. The PPDs may not perfectly approximate observed differences due to rounding to the nearest integer.

As can be seen in Table 13, some officers have significantly larger proportions of stops and arrests involving Black civilians relative to their peers who conducted stops and arrests in similar times, places, and contexts. However, given the concerns with conducting the same benchmark test a few hundred or perhaps several thousand times—see Equation (6), I employ a multiple-hypothesis testing correction by controlling for the local false discovery rate. This will shed light on which of these z-statistics remains statistically significant after controlling for an increased risk of Type-I error, which allows me to be more certain about the accuracy of the internal benchmark analyses.

False Discovery Rate Results

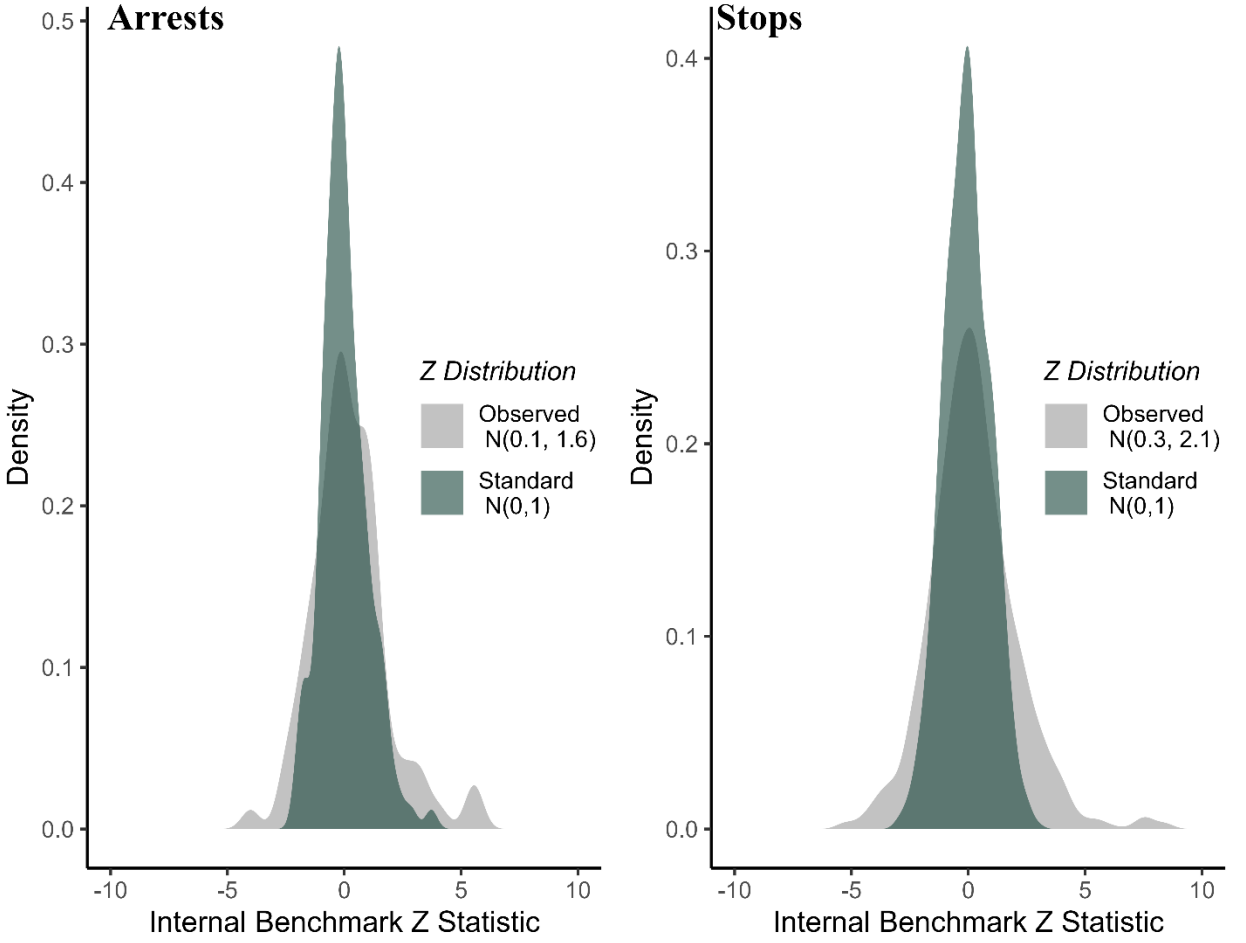
Arrests

Among the 139 officers in question with matched arrest benchmarks, approximately 20 arresting officers (14%) would have been deemed an outlier based on the absolute value of their z-statistic being at least 1.96, or what is traditionally referred to as the 95% confidence level in a standard two-tailed null hypothesis test. However, there is an increased risk of Type-I error given the multiple hypothesis testing situation. Accordingly, I reduce the risk of falsely identifying an

officer as arresting far more (or fewer) Black civilians relative to their peers by re-estimating the statistical significance of these z-statistics based on the empirically derived null distribution of all officers' z-statistics in the arresting officer dataset.

Figure 10 visualizes the z-statistic distributions, where the standard null $N(0,1)$ is reflected by the dark green shaded region and the observed null $N(0.1,1.6)$ is reflected by the light grey shaded region. Each shaded region reflects the density distribution of their z-statistics. The observed null has a slightly wider distribution than standard normal, which bears important implications for false positive identification. Recall that to determine whether an officer exceeds their benchmark given their z-statistic, I must first determine what the probability is that they do not exceed their benchmark (i.e., FDR), which I obtain through the observed distribution of the z-statistics in Figure 10. Then, I can determine what the probability is that an officer is an outlier given their z-statistic and compare this to their FDR.

Figure 10. Observed and Standard Normal Distributions of Z-Statistics



In both stops and arrests, officers with z-statistics that fall just to the right of the 97.5th percentile (or just to the left of the 2.5th percentile) on a standard normal distribution would be designated as an outlier based on the standard normal distribution. Yet, as can be seen by the density distribution of the observed z-statistics, not all officers would fall beyond the same percentiles given how many more officers have z-statistics in that region. Comparing officer z-statistics to the observed distribution will provide a more conservative estimate of officers who are potential outliers.

According to Ridgeway and MacDonald (2009), any officer who has a higher probability of being flagged as an outlier than not being one (FDR < .50) would be considered to have a

statistically significant z-statistic—given the distribution of observed z-statistics across all officers with suitable benchmarks. Accordingly, those with $z > 3.64$ represent 4 of the 139 arresting officers who have a higher probability of being an outlier given their z-statistic than not being one. These officers represent the right side of the observed null distribution of z-statistics, thereby indicating that they arrested a greater proportion of Black civilians relative to their peers in similar times, places, and contexts. On the other side of that distribution, those officers with $z < -3.56$ represent 3 of the 139 arresting officers, of whom stopped far fewer Black civilians compared to their peers. In total, 7 officers were identified as at-risk for engaging in racially-disparate arrests.

To expand on the work of Ridgeway and MacDonald (2009), I show how the results can be disaggregated even further by how much Type-I error I am willing to accept in the internal benchmark analyses. Accordingly, in Table 14 the results of how many officers are flagged as outliers and are reported by direction of estimated disparity and by false discovery rate. As the false discovery rate tolerance decreases (smaller Type-I error), the number of flagged officers decreases.

Table 14. Estimated Number of Officers with Significant Disparity Estimates by Direction of Disparity and False Discovery Rate Tolerance

False Discovery	Arrests			Stops		
Rate Tolerance	Positive	Negative	Total	Positive	Negative	Total
FDR ≤ 50%	4	3	7	134	31	165
FDR ≤ 25%	3	3	6	81	11	92
FDR ≤ 5%	2	0	2	50	5	55

Notes: FDR represents the false discovery rate

Stops

Turning now to the sample of stopping officers with suitable benchmarks (N = 3,010), results indicate that 712 of them had a statistically significant doubly robust estimate of racial disparity at the 95% confidence level using the traditional null hypothesis test. Much like the arrest behavior, the observed distribution of z-statistics for stopping officers is much wider than the

standard normal distribution (See Figure 10). When examining only those officers whose chances of being flagged as an outlier exceeded their chances of not being one, the results indicate only 165 officers were at-risk for engaging in disparate stops. Officers with $z > 3.91$ represent the 134 officers who had a higher probability of being an outlier given their statistic than not being one. These officers represent 81% of outliers and were flagged as such due to stopping a greater proportion of Black civilians relative to their peers who worked in similar circumstances as they did. The remaining 31 officers with $z < -4.25$ represent the 19% of officers who stopped significantly fewer Black civilians relative to their peers.

Summary

There are two initial takeaways from these findings worth noting. First, officers who had statistically significant disparity estimates in the benchmark analysis represent roughly 5% of all arresting officers and stopping officers in the analyzed samples of data. This suggests that racial disparities in officer behavior are concentrated among a small portion of police officers across the agency. Although concentrations of criminogenic behavior and police misconduct have been reported elsewhere (Christopher, 1991; Gottfredson & Hirschi, 1990; Wolfgang et al., 1972), such a finding has been largely unreported in the etiology of racial disparities in police behavior.

After controlling for the risk of false discovery, the racial disparity identification procedure employed in this study revealed considerably fewer outliers than previously suggested based on the standard null hypothesis testing procedure. Specifically, the officers I identified as at risk for racially-disparate arrest behavior (N = 7) represent only 23% of all officers who had significant z-statistics for their internal benchmark analysis based on standard null hypothesis testing procedures. Likewise, I identified only 44% of officers (N = 165) who originally had significant z-statistics for their traffic stop internal benchmark. As shown in Table 15, the number of officers

flagged as outliers only decreases as I further safeguard against the risk of false positive identification.

This highlights the importance of accounting for the risk of false positive identification in racial disparity analyses at a micro level. Failing to do so may lead to drastically different conclusions about who potentially engaged in racially-disparate police behavior. Conducting conservative analyses is thus important not just for reducing statistical bias, but for preserving an organizationally-just climate through procedurally fair evaluation practices (Wolfe & Lawson, 2021). More broadly, being certain about who may be at risk and knowing the contextual factors associated with why represents a key strength of the internal benchmarking procedure and offers an important source of information that agency leaders can draw on to develop officer-specific and problem-oriented reforms.

It should be noted here briefly that there are some potential drawbacks to the internal benchmarking procedure. Most notably, it is a data-driven strategy that requires an ample number of stops/arrests per officer. This was lacking among many CPD officers. The implications of having a reduced sample will hinder the generalizability of the procedure's findings to the agency. The practical utility of deploying this procedure in smaller agencies thus becomes a concern given many smaller agencies do not face the same extent of crime problems and officer demand that CPD faces. The overall theoretical and practical implications of these findings will be discussed in more detail in the concluding chapter of the dissertation.

Descriptive Statistics of Outlier Officers

Correlates of Outlier Officers

Before analyzing policy-relevant questions related to officer diversification, a key question relevant to theoretical and political discourse about race and gender in policing is to what extent

are these characteristics associated with officers who engage in racially-disparate behavior? For example, are males more likely than females to be at risk for racial profiling? Some evidence suggests that females are less likely to stop and arrest Black civilians, which could mean they are less likely to racially profile in these behaviors (Ba et al., 2021). However, previous studies often fail to account for important confounding factors (and the interactions between them) that could explain racial differences in their behavior, such as where in their beat assignments they contact civilians, at what time of day they make these contacts, and in what socioeconomic contexts.

Fortunately, the main advantage of conducting the internal benchmarking procedure is that it provides an internally valid and statistically unbiased estimate of racial disparity for each officer in the agency by accounting for where, when, and under what circumstances they make their stops and arrests of Black civilians. What remains unexplained after conducting these benchmark analyses is whether the officers who have a statistically significant estimate of racial disparity differ on any key demographic characteristics from those who did not have significant disparity estimates. Accordingly, in Table 15, I compare the demographic composition of officers who have statistically significant estimates of racial disparity with those who do not to see whether there are any differences in their individual characteristics.

Table 15. Comparison of Officer Characteristics Between Officers with and without Racially Disparate Arrests and Stops

Arrest Data		At-Risk Officers (N = 7)	Not At-Risk Officers (N = 132)	*t* (X^2)	*p*
Race	Black	0%	9%	(1.07)	0.90
	Hispanic	29%	30%		
	Other	0%	0%		
	White	71%	58%		
Gender	Female	14%	12%	(0.06)	1.00
	Male	86%	88%		
Spanish Speaking	No	86%	89%	(0.09)	1.00
	Yes	14%	11%		
Age	Mean	41	40	0.71	0.44
Total Arrests	Mean	92	76	0.81	0.34

Stop Data		At-Risk Officers (N = 165)	Not At-Risk Officers (N = 2,845)	*t* (X^2)	*p*
Race	Black	22%	20%	(4.42)	0.34
	Hispanic	24%	25%		
	Other	4%	5%		
	White	50%	50%		
Gender	Female	21%	22%	(0.29)	0.63
	Male	79%	78%		
Spanish Speaking	No	88%	87%	(0.51)	0.48
	Yes	12%	13%		
Age	Mean	40	41	2.43	0.02
Total Stops	Mean	203	178	1.84	0.05

Notes: All independent sample t-tests are estimated based on 10,000 bootstrapped replications. P values represent bias-corrected estimates based on bootstrapping procedures. All Pearson X^2 tests are estimated using p-values computed by 10,000 Monte Carlo simulations. All two-sample t-test p-values are estimated with 10,000 bootstrap replications.

Results from Table 15 reveal four main findings. First, there were no statistically significant differences in the racial distribution of arresting officers who have significant racial disparity estimates compared to those who do not. On a related note, the racial and ethnic distributions are quite similar and not significantly different between officers who engage in disparate stop behavior compared to those who did not engage in disparate stops.

Another interesting finding reported in Table 16 is that the composition of males and females is quite similar between the sample of officers with and without racial disparities in their stops and arrests. What this suggests more generally is that gender does not predict whether officers have racial disparities in their behavior.

When examining other officers' demographic information, results indicate that officers with racial disparities in their stops are about a year younger than those who do not have disparities. While this difference is quite small, it does correspond with other studies that have found younger officers tend to have more complaints and allegations of misconduct against them (Kane & White, 2009; Wolfe & Piquero, 2011). The findings also reveal that officers who were deemed as having significant racial disparities in their arrests have a similar total number of arrests as those without disparities. However, those with stop disparities generally conducted more stops than those who did not, which might align with past literature suggesting aggressive enforcers may be more inclined to engage in discretionary and perhaps racially-disparate behavior (Skolnick, 1966).

Behavioral Invariance of Racial Disparities

One of the primary motivations behind analyzing multiple types of behavior in a racial disparity analysis is that officers may have different "tastes" for racially-disparate policing, which would be reflected in the different forms of behavior they engage in. Theoretical discourse on the etiology of racial discrimination in policing has largely assumed that such behavior would be

concentrated among a few officers who harbor deep-rooted racial prejudices towards a minority group of people. This implies that officers who engage in racially-disparate policing must do so across all their behaviors. Neither theoretical nor empirical research has yet to consider whether discrimination is behaviorally invariant at an individual level—rather it has been merely implied (see Thomaskovic-Devey et al., 2004). The question then becomes whether we should expect officers to exhibit racial disparities across multiple types of their behavior; that is, are racial disparities in the CPD behaviorally invariant?

To answer this question, I calculated the percentage of officers who had significant and not significant disparity estimates for their arrest and stop behaviors. Among 3,043 officers, 3% of officers had racial disparities in their arrest *and or* stop behavior. Among these 106 officers, 8% of officers (N = 8) had significant disparity estimates only among their stops, about 1% (N = 1) had a significant disparity estimate for only their arrests, and almost 1% of officers (N = 2) had significant disparity estimates among their stops *and* arrests.

It should be noted that these findings represent a very small subset of officers to begin with. These findings should be interpreted cautiously and replicated frequently in future research to determine the consistency of these findings. What can be inferred from these findings is, however, that racial disparities were not invariant across police behaviors in the CPD. Some officers have disparities in their stop behavior while others do in their arrest behavior. In the CPD, concern about racial disparities should not be isolated to any single behavior, otherwise, those officers identified as at risk for racially-disparate policing may not reflect the entire universe of officers who were perpetuating racial inequalities in the agency.

Officer Diversification Results

An unsolved question in current police reform debates is whether increasing racial and

gender diversity in a police agency can reduce racial disparities in police outcomes. Previous results suggest that race may play a role in predicting officers' risk for engaging in racially-disparate police behavior, but gender does not predict such behavior. However, the question remains whether under-represented minority officers engage in more racially equitable policing than their male and White colleagues.

Facilitating this empirical inquiry is a three-step process. First, I constructed four internal benchmarks for each outlier officer—each benchmark differs on one key demographic characteristic (all Black, White, female, or male officers, respectively). Officers who make stops and arrests within these benchmarks are not outliers and thus represent the majority of officers who do not engage in systematically disparate police behavior. Using this information, I then generated a doubly robust estimate of racial disparity for each of the four benchmarks for each outlier officer, which is measured as a percentage point difference between an outlier's racial composition of stops/arrests relative to their benchmark. Finally, I estimated the difference in the percent-point difference between the Black officers' benchmark and the White officers' benchmark across all outlier officers and did the same for female and male officer benchmarks.

Taking the difference in the distribution of these disparity estimates between benchmarks and across all outlier officers will indicate whether the observed disparities in arrests or stops are greater (or smaller) when comparing all Black officers that do not engage in racially-disparate behavior to all White officers that also do not engage in racially-disparate behavior, and when comparing all female officers that do not engage in racially-disparate behavior to all male officers that do not engage in racially-disparate behavior. If, for example, disparities are larger for the female benchmarks than the male benchmarks, taking this difference in the distribution of their disparities would indicate whether females engage in more racially equitable police behavior than

their male colleagues who work in the same times, places, and contexts as they do. In other words, any marginal difference between benchmark disparity estimates is presumed to be a marginal gain for racially equitable policing because neither group of officers engaged in disparate policing but may stop or arrest a smaller proportion of Black civilians—a presumed "win" for diversification advocates.

Table 16 presents the results of the officer diversity analyses, where disparity estimates are presented as the average percentage point difference (APPD) in the proportion of stops/arrests involving Black civilians for outlier officers relative to their internal benchmarks, which are comprised of only Black officers and only White officers who were not deemed as outliers based on the previous internal benchmark analyses. In the top panel, I consider only outlier officers who were originally flagged as arresting and stopping a greater proportion of Black civilians relative to their peers based on the standard FDR cutoff.

Table 16. Disparity Estimates by Race and Direction of Original Disparity Estimate

Originally Positive Disparity	APPD			95% CI	
	Black	White	*p*	Lower	Upper
Arrests					
FDR ≤ 50%	13	14	0.89	-9	8
FDR ≤ 25%	10	15	0.33	-12	2
FDR ≤ 5%	11	14	0.75	-14	8
Stops					
FDR ≤ 50%	6	7	0.01	-3	-1
FDR ≤ 25%	5	6	0.13	-3	0
FDR ≤ 5%	5	6	0.44	-3	1
Originally Negative Disparity	**APPD**			**95% CI**	
	Black	White	*p*	Lower	Upper
Arrests					
FDR ≤ 50%	-28	-9	0.07	-34	-9
FDR ≤ 25%	-28	-9	0.08	-34	-9
FDR ≤ 5%	-	-	-	-	-
Stops					
FDR ≤ 50%	-18	-13	0.01	-10	-1
FDR ≤ 25%	-20	-14	0.15	-13	2
FDR ≤ 5%	-28	-13	0.16	-24	-3

Notes: (APPD) represents the average percentage point difference in the racial composition of stops/arrests for outlier officers relative to their internal benchmarks. False discovery rates (FDR) represent the probability that an officer exceeds their benchmark given the z-statistic they were assigned—smaller cutoffs correspond to a smaller number of "bad apple" officers in each sample. Differences were estimated using a two-sample t-test with 10,000 bootstrap replications. (-) indicate there were no estimates due to there being no officers with negative disparity estimates at that indicated FDR level.

For example, in examining the first row of the top panel, the results show that outlier officers have, on average, a larger proportion of arrests involving Black drivers by about 13 percentage points (AAPD-Black = 13) when compared to their Black colleagues who do not engage in disparate arrest behavior. In contrast, there is a 14-percentage point difference (AAPD-White = 14) in the proportion of arrests involving Black civilians when comparing outlier officers to their White colleagues. At face value, this might suggest that White officers engage in more racially equitable arrests when compared to Black officers. However, taking the difference in these yields a 1 percent-point difference, with estimates ranging between -9 and 8 at the 95% confidence

level. This null finding holds across FDR cutoffs for arresting officers when considering outliers who arrested a greater proportion of Black civilians relative to their benchmark.

When considering positive disparities for stopping officers, a different set of findings emerged. Whereas outlier officers have a 6-percentage point difference (AAPD-Black = 6) in the proportion of stops involving Black civilians when compared to their Black colleagues, there is a 7-percentage point difference (AAPD-White = 7) in the proportion of arrests involving Black civilians when comparing outlier officers to their White colleagues. In other words, White officers were stopping a slightly smaller proportion of Black citizens relative to their Black colleagues who work in the same times, places, and circumstances. Unlike arrests, the difference in these AAPDs was statistically significant, with estimates ranging between -3 and -1. More practically, this suggests that White officers may engage in slightly more racially equitable stops when compared to their Black colleagues. Importantly, this finding does not hold when trimming down the number of outliers to reduce the risk of falsely attributing officers as engaging in racially-disparate behavior.

The bottom panel of Table 16 considers only outlier officers who were flagged as arresting/stopping a smaller proportion of Black civilians relative to their peers. AAPD estimates are measured negatively to reflect this underrepresentation. When looking at the first row of the bottom panel, the results show that outlier officers have, on average, a smaller proportion of arrests involving Black drivers by about 28 percentage points (AAPD-Black = -28) when compared to their Black colleagues who do not engage in disparate arrest behavior. In contrast, there is a 9-percentage point difference (AAPD-White = -9) in the proportion of arrests involving Black civilians when comparing outlier officers to their White colleagues. This suggests that White officers conduct more racially equitable arrests when compared to their Black colleagues.

However, with the small number of officers within this subgroup of outliers, these findings should be interpreted cautiously.

The results for stops remain consistent across the direction of the observed racial disparity. As seen in the bottom panel of Table 16, whereas outlier officers have an 18-percentage point difference (AAPD-Black = -18) in the proportion of stops involving Black civilians when compared to their Black colleagues, there is a 13-percentage point difference (AAPD-White = -13) in the proportion of stops involving Black civilians when comparing outlier officers to their White colleagues. The difference between these estimates varied between 10 and 1 at the 95% confidence level, suggesting, again, that White officers may engage in more racially equitable stops compared to Black officers.

Table 17 presents the results when internal benchmarks comprise only female officers and only male officers. More specifically, the first two columns represent the AAPD when comparing outlier officers to internal benchmarks comprised of stops and arrests made by only female officers and by only male officers who did not engage in racially-disparate behavior according to previous analyses. Results from both panels indicate that the disparity estimates generated from the female benchmarks were slightly larger than for the male benchmarks when considering outlier officers who conducted more arrests and stops involving Black civilians relative to their peers. These differences are, however, not statistically significant across all FDR cutoffs. This suggests that female officers and male officers make a similar proportion of arrests involving Black civilians when they are situated in the same times, places, and contexts. This finding holds when considering outlier officers who conducted a smaller fraction of arrests/stops involving Black civilians relative to their benchmark as well.

Table 17. Disparity Estimates by Gender and Direction of Original Disparity Estimate

Originally Positive Disparity	APPD			95% CI	
	Female	Male	*p*	Lower	Upper
Arrests					
FDR ≤ 50%	12	11	0.83	-7	11
FDR ≤ 25%	11	10	0.52	-7	16
FDR ≤ 5%	15	10	0.51	-7	17
Stops					
FDR ≤ 50%	7	7	0.84	-1	2
FDR ≤ 25%	7	6	0.72	-2	2
FDR ≤ 5%	7	6	0.79	-2	3
Originally Negative Disparity	**APPD**			**95% CI**	
	Female	Male	*p*	Lower	Upper
Arrests					
FDR ≤ 50%	-24	-11	0.60	-29	9
FDR ≤ 25%	-24	-11	0.60	-29	8
FDR ≤ 5%	-	-	-	-	-
Stops					
FDR ≤ 50%	-15	-13	0.15	-5	1
FDR ≤ 25%	-19	-15	0.15	-9	1
FDR ≤ 5%	-18	-14	0.46	-12	5

Notes: APPD represents the average percentage point difference in the racial composition of stops/arrests for outlier officers relative to their internal benchmarks. False discovery rates (FDR) represent the probability that an officer exceeds their benchmark given the z-statistic they were assigned—smaller cutoffs correspond to a smaller number of "bad apple" officers in each sample. Differences were estimated using a two-sample t-test with 10,000 bootstrap replications. (-) indicate there were no estimates due to there being no officers with negative disparity estimates at that indicated FDR level.

Summary

The results from the diversification policy analyses bear two key findings. First, officers' race may be inconsistently related to the racial equity of their police work. Previous research shows an inconsistent relationship between officers' race and engaging in racially-disparate behavior (Ba et al., 2021; Brown & Frank, 2006; Headley & Wright, 2020). Findings from this study support this, albeit they are highly sensitive to the level of risk regarding false positive identification of outlier officers (i.e., FDR cutoff). Given the stark under-representation of Black officers in the analyzable sample of data, it could be that these results are idiosyncratic to a subset of Black officers in the agency who arrest and stop more Chicagoans than most of their Black colleagues.

These findings should thus be interpreted within the context of the sample, not necessarily the agency as a whole. Second, there were no gender differences in the racial equity of officers' arrests when comparing male and female colleagues who make stops and arrests in the same set of circumstances. These results should be interpreted cautiously as they reflect behavior among a very small number of female CPD officers who are particularly active compared to their fellow female colleagues.

Police Beat Assignment Results

Results from previous analyses suggest that officers' race and gender are inconsistent risk factors for predicting which officers engage in racially-disparate behavior. Given the potential concern with the consistency of those results, the next set of analyses test another potential avenue for police reform that draws on a situational crime prevention framework to see whether change in opportunity structures can alter racially-disparate police behavior. I analyze how the racial composition of arrests/stops made by an officer who is at risk for engaging in racially-disparate police behavior (bad apple) differs from that of their peers when working in the same police beats on the same days over the four years.

As one might expect, simply because an officer was found to be an outlier in the original analysis over four years, does not mean they will always engage in racially-disparate behavior every day during every assignment. Accordingly, the focus of this analysis is to see whether on days when they do engage in that behavior, it is more likely to occur when they are assigned to police beats where the residential population has higher concentrations of Black civilians. The logic behind the analysis is that the risk of perpetuating disparities can be potentially mitigated by inhibiting the chances of these officers from arresting/stopping Black civilians in the first place.

Other theoretically relevant variables were included in the models to ensure these estimates

were as precise and unbiased as possible. For example, I included the level of violent crime in a police beat, the number of other officers that were deemed as outliers over the four years, and temporal factors that help explain day-to-day, month-to-month, and year-to-year variation in disparities. I also included a categorical variable indicating which officer is being analyzed in the model to account for individual variation in the propensity to engage in disparate behavior that may be driven by unobserved officer characteristics.

The outcome of this analysis (racial disparity) is categorical in nature given that most observed differences in the racial composition of arrests/stops between outlier officers and their peers were zero, while a small fraction of observations involved outlier officers having arrested/stopped a greater and lesser proportion of Black civilians than their peers on any given day in a given beat. Accordingly, when interpreting the results of the logistic regression analyses, all estimates are based on a comparison between no disparity to either a positive or negative disparity. Whereas positive disparity indicates an outlier officer arrested/stopped a greater proportion of Black civilians than their peers in the same beat and day, negative disparity indicates they stopped a smaller proportion of Black civilians than their peers. Results from these analyses are presented in Table 18.

Table 18. Logistic Regression Results of Officer Disparate Behavior by Beat Composition

Variable	Arrests (N = 1,357)		Stops (N = 46,438)	
	No Disparity Compared to…		No Disparity Compared to…	
	Negative Disparity (N = 64)	Positive Disparity (N = 96)	Negative Disparity (N = 12,499)	Positive Disparity (N = 4,130)
Beat Measures				
% Black	18.79**	116.18**	3.40**	2.42**
	(0.86)	(1.00)	(0.26)	(0.21)
Violent Crime	1.04	1.33	1.04	0.94*
	(0.22)	(0.23)	(0.04)	(0.03)
# of Bad Apples	-	5.18**	0.78*	1.06
	-	(0.45)	(0.11)	(0.11)

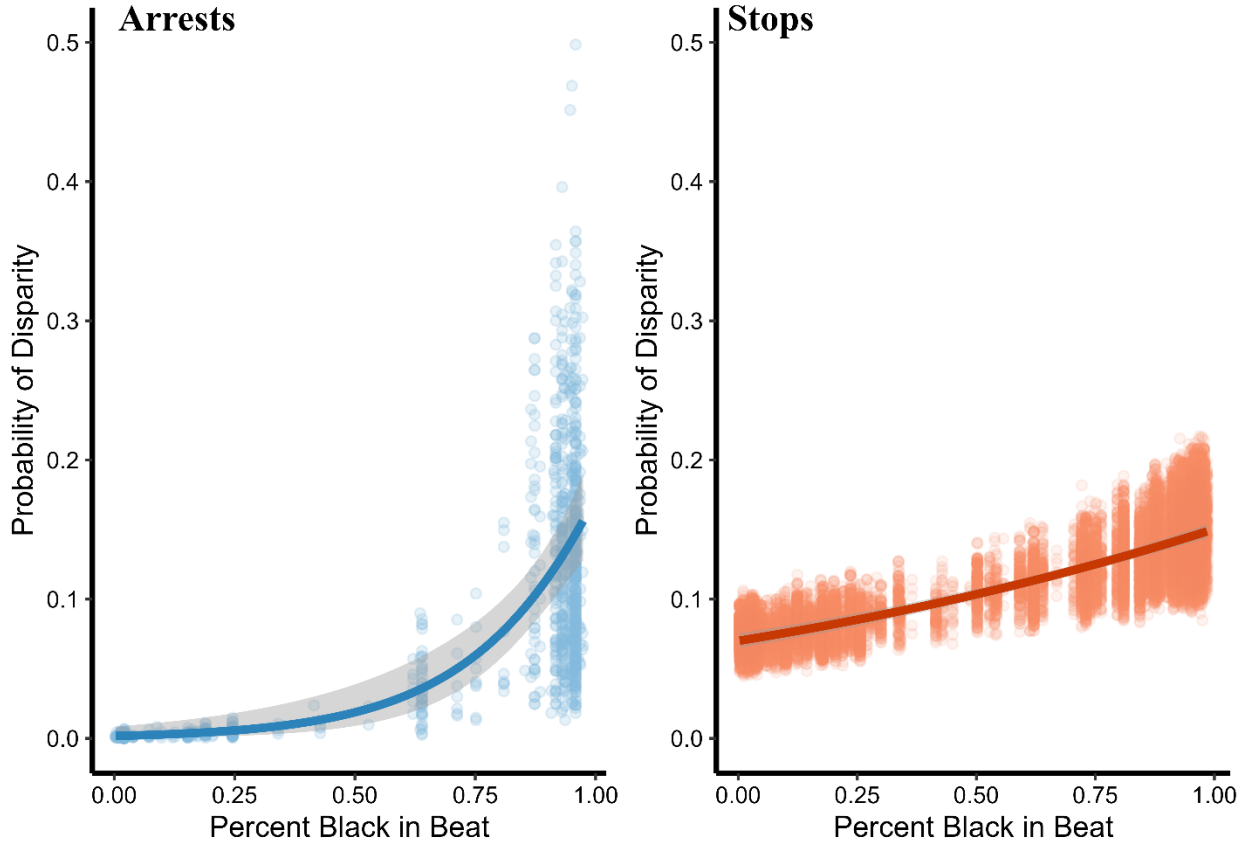
Notes: *$p \leq 0.05$; ** $p \leq 0.01$; Results measured as Odds Ratios. Standard errors are clustered at the beat level and noted in parentheses. N corresponds to the number of officer×day×beat observations for each dataset. Positive disparity indicates bad apple officers arrested/stopped proportionally more Black civilians than their peers in the same beat and day. Negative disparity indicates bad apple officers arrested/stopped proportionally less Black civilians than their peers in the same beat and day. All models include day of week, month of year, year, and officer-in-question fixed effects. (-) indicate there were no estimates due to there being no data for that variable.

Results from Table 18 reveal three main findings. First, the findings indicate that when comparing the racial composition of stops and arrests made by a 'bad apple' officer and their peers on the same day and in the same police beat, the odds of observing a racial disparity in that officer's arrests/stops is significantly higher when they work in police beats that have larger proportions of the residential population that is Black. In other words, officers who are at risk for engaging in racially-disparate arrest and stop behavior are more likely to engage in that behavior when they work in jurisdictions with a higher proportion of Black residents.

To better understand this finding, Figure 11 displays the predicted probability of there being an observed positive disparity in an outlier officer's arrests/stops in police beat *k* on day *t.* All predicted probabilities are about the percentage of a police beat's residential population that is Black, holding all other variables in Equation (9) at their means. While the predicted probability is displayed as a best-fit line with 95% confidence bands, also displayed are the estimated

probabilities themselves (shown as dots). Results indicate that the probability of an outlier engaging in a disproportionately higher number of stops and arrests in a given day, for a given beat, for a given outlier officer increases considerably as the residential population of that beat has higher concentrations of Black residents.

Figure 11. Predicted Probability of Disparity



The results for arrests should be interpreted cautiously in Table 18. These estimates were based on seven officers' arrests, which led to few instances of any disparity, and thus a high degree of inconsistency in the parameter estimates. Moreover, because there were so few officers deemed outliers to begin with for the arrest data, there were no instances in which multiple outlier officers were working in the same beat on the same date, which meant that there was no way to estimate the effect of having more than one bad apple officer in the same beat.

In addition, there was never an instance of a positive disparity where the majority of the

residential population was not Black to begin with (that is, greater than 50% of the population in a police beat). There were only four instances of a negative arrest disparity when an outlier officer worked in a police beat where the population was not predominantly Black as well. In other words, arrest disparities for these officers almost entirely occurred in police beats where the residential population was predominantly Black. This is perhaps surprising given that these outlier officers were assigned to police beats where the residential population was predominantly Black less than half the time. Put simply, the sheer infrequency of disparities coupled with the fact that they only ever existed in predominantly Black police beats meant that the racial composition of a police beat largely explained the disparities perpetuated by outlier officers. Given these data patterns and potential limitations, the large odds ratios reported for positive *and* negative arrest disparities in Table 18 may be reality, or they may be an artifact of the available data.

Another key finding from Table 18 is that the probability of officers stopping fewer Black civilians than their peers is greater when outlier officers work in police beats with more Black civilians. This is counterintuitive given the theoretical motivations behind the analysis. If anything, one would anticipate negative disparities to be less likely to occur in settings where there are fewer Black citizens living in a police beat. One potential explanation is that some officers may be cognizant of the risks of engaging in racially-disparate behavior when working in higher-risk contexts, thereby motivating them to stop and arrest fewer Black civilians than their peers. Unfortunately, without quantitative or qualitative information based on CPD officer accounts, the nature of this result is not well understood in the current study setting.

One interesting finding worth noting is that the chances of an officer stopping more Black civilians relative to their peers decreases as they work in more violent beats—independent of the racial composition of the beat itself. One potential explanation for this is that officers may be more

inclined to conduct strategic stops and are more selective in whom they stop when working in violent areas. That is, the threshold for stopping citizens is simply higher and perhaps less influenced by extra-legal factors such as one's race. Their decision-making in these more dangerous beats is thus driven by a focus on officer safety, crime prevention, and or violence reduction as opposed to implicit stereotypes or cognitive shortcuts.

The final key takeaway from Table 18 is that the odds of there being a positive disparity in arrests for an outlier officer *i* in police beat *k* on day *t* is significantly greater when multiple officers are working in that beat on that day who were previously identified as being at risk for engaging in racially-disparate police behavior over the four years. In contrast, the odds of there being a negative disparity in stops is far less likely as there are more bad apple officers working in the same beat on the same day. This suggests that having more deviant peers in a beat may lead to a higher chance of an outlier officer arresting more Black civilians relative to their non-deviant peers in the same police beat and a lower chance of an outlier officer stopping fewer Black civilians relative to their non-deviant peers.

Table 19 reports the results of the beat assignment analysis when restricting attention to only those officers whose probability of being an outlier exceeded 75% and 95%, respectively. As seen in the top and bottom panels, the results are consistent with the previous analyses when examining stop outcomes. However, when examining stop outcomes, there was no longer a relationship between an outlier officer stopping more Black citizens relative to their peers and the concentration of Black civilians living in that beat. This suggests that the findings may be sensitive to how certain we are of who engaged in racially-disparate police behavior. The implications of these findings will be discussed in the conclusion of this study. These findings should be interpreted cautiously, especially for the arrest data, given how few observations there are in both

positive and negative disparity outcomes as the FDR cutoffs shrink.

Table 19. Logistic Regression Results of Officer Disparate Behavior by Beat Composition with 25% and 5% False Discovery Rate Cutoffs

FDR ≤ 25%				
	Arrests (N = 1,362)		Stops (N = 24,860)	
	No Disparity Compared to…		No Disparity Compared to…	
Variable	Negative Disparity (N = 64)	Positive Disparity (N = 95)	Negative Disparity (N = 7,300)	Positive Disparity (N = 2,183)
Beat Measures				
% Black	24.69**	142.52**	3.54**	1.34
	(0.88)	(1.03)	(0.19)	(0.30)
Violent Crime	1.01	1.32	1.01	0.91*
	(0.22)	(0.23)	(0.03)	(0.05)
# of Bad Apples	-	5.43**	0.76**	1.14
	-	(0.46)	(0.06)	(0.14)
FDR ≤ 5%				
	Arrests (N = 643)		Stops (N = 16,179)	
	No Disparity Compared to…		No Disparity Compared to…	
Variable	Negative Disparity (N = 54)	Positive Disparity (N = 70)	Negative Disparity (N = 5,080)	Positive Disparity (N = 1,249)
% Black	1.06	281.89*	3.93**	1.09
	(1.82)	(2.55)	(0.24)	(0.38)
Violent Crime	1.00	0.87	1.01	0.89
	(0.21)	(0.29)	(0.03)	(0.07)
# of Bad Apples	-	-	0.71**	1.05
	-	-	(0.06)	(0.13)

Notes: *$p \leq 0.05$; ** $p \leq 0.01$; Results measured as Odds Ratios. Standard errors are clustered at the beat level and noted in parentheses. N corresponds to the number of officer×day×beat observations for each dataset. Positive disparity indicates bad apple officers arrested/stopped proportionally more Black civilians than their peers in the same beat and day. Negative disparity indicates bad apple officers arrested/stopped proportionally less Black civilians than their peers in the same beat and day. All models include day of week, month of year, year, and officer-in-question fixed effects. (-) indicate there were no estimates due to there being no data for that variable at that the indicated FDR level.

CHAPTER 5: CONCLUSION

Understanding the prevalence of racial inequalities in policing is a key focus of criminological research and bears important implications for public health and public safety among people of color. As such, many scholars have directed their attention toward identifying when racial disparities in police behavior may exist, and what may give rise to these disparities in the first place (Knowles, Persico, & Todd, 2001; Pierson et al., 2021; Rojek, Rosenfeld, & Decker, 2012). Lacking, however, is consistent evidence that documents the effectiveness of policy innovations and police reforms to mitigate the consequences of racially-disparate police behavior.

To instigate evidence-based police reform, police command staff and policymakers must consider what level of analysis their proposed reforms will be deployed as this will help achieve their intended benefits (Lum & Koper, 2017). This requires proper specification of how to *identify* and *remediate* racial disparities in police behavior at the same level of analysis. Integrating a focus on identification and remediation is critical for catering innovations to police reform that reflect the level at which inequalities exist. Doing so may provide police leaders with an ideal evidence-based framework to understand and address racial disparities in policing.

Present Study

The present study sought to make three contributions to the literature on racial inequalities in policing. The initial contribution of this study is to determine if racial inequalities in police behavior are perpetrated by a subset of officers in a police agency. By analyzing data from police officer shift assignments, stops, and arrests in the Chicago Police Department between 2012-2015, this study specifically seeks to identify which officers disproportionately stopped and arrested Black civilians after accounting for when, where, and under what circumstances they conducted their daily police work.

Research has shown that police officer misconduct can be concentrated among a few officers in the agency (Chalfin & Kaplan, 2021; Christopher, 1991), and this can be true for racial disparities in traffic stop behavior as well (Ridgeway & MacDonald, 2009). However, the evidence in favor of the latter finding is scant and relies on limited contextual information as to where, when, and under what circumstances police officers work—all of which may help better explain their disparate behavior. Drawing on a large source of data across a variety of contexts in one of the largest and most diverse police departments in America, this study sought to shed new insights into the study of racial disparities in policing.

The second contribution of this study is to test a series of policy-relevant questions that may give rise to potential policy interventions based on the scope of racial disparities observed in an agency. The first policy-relevant question I explored was whether officers with under-represented racial and gender minority identities were more likely to promote racially equitable behavior than their White and male colleagues. In drawing on theories of representative bureaucracy, officers may be better equipped to engage in positive police-citizen interactions if they can understand and address civilians' life situations. As others have suggested, this may be more likely to occur when officers reflect the demographic composition of their constituents.

Another policy-relevant question I explored was whether the residential composition of a police beat was related to the probability of officers engaging in racially disparate behavior while restricting attention to those officers who were at the most risk for engaging in racially disparate behavior in the first place. Building on a situational crime prevention framework, I proposed that police beats with higher concentrations of Black residents may reflect high-opportunity contexts for racially disparate police behavior. If there are officers who are at an increased risk of engaging in racially disparate police behavior, then having these individuals in predominantly Black police

beats could amplify the chances that they perpetuate disparities. Accordingly, I assessed whether outlier officers who worked in settings with more opportunities to contact Black citizens based on the residential composition of the police beat an officer worked in was related to a higher probability that they engaged in racially disparate behavior.

Lastly, this study contributes to the ongoing body of research that seeks to understand how best to estimate racial disparities in police behavior by offering new insights into a micro-level approach. While research on internal benchmarking procedures for racial disparities is not new, this study shows that there are several relevant factors often absent in previous internal benchmark analyses due to data limitations. I also show how findings may change based on the risk of Type-I error that an agency is willing to accept when conducting internal benchmark analyses is large police agencies, thereby paving the way for more transparent and theoretically informed approaches in the future.

Summary of Findings

Several findings emerged from the analyses that warrant further discussion concerning past research. The first set of analyses explored whether racial inequalities in police behavior were perpetrated by some but not all officers in the agency. In alignment with past research (Ridgeway & MacDonald, 2009), the results revealed that racial disparities in stops and arrests were perpetrated by some but not all officers in the agency. More specifically, 5% of stopping officers and arresting officers were found to have engaged in racially-disparate behavior, while most other officers did not. The theoretical implications of this finding are that much like crime itself, racial inequalities in police behavior can be concentrated at a micro level. This is one of the few studies thus far to report such an empirical finding in the study of racial inequalities in policing (Ridgeway & MacDonald, 2009). This finding also has practical relevance for Sherman's (2007) power few

hypothesis as it appears that racial disparities in police behavior may be well-suited for targeted reform efforts.

As the later results reveal, scholars and police practitioners should consider the possibility that racial disparities can vary by the behavior in question. When restricting attention to the 106 officers from which I could generate adequate internal benchmarks for their stop *and* arrest behavior, the results showed that eight officers had significant disparity estimates only among their stops, one officer had a significant disparity estimate for only their arrests, and two officers had significant disparity estimates among their stops *and* arrests. This provides preliminary evidence that, from a theoretical standpoint, racial disparities in CPD may not be behaviorally invariant.

There are, however, two competing schools of thought on the behavioral invariance of racial disparities in police behavior—assuming such behavior could be broadly defined as a form of organizational deviance. On one hand, adopting a generality of deviance perspective would lead some to argue that there is truly no behavioral invariance in the disparate behaviors observed by a small number of CPD officers. Rather, what was observed can be attributed to different opportunity structures that help explain why some officers engaged in racially disparate stops in the CPD and others engaged in disparate arrests. Indeed, while having low levels of self-control tends to be a consistent predictor of juvenile delinquency and crime, such behavior can cease to exist across individuals with low self-control because there lacks adequate opportunity structures to offend in the first place (Gottfredson & Hirschi, 1990; Hay & Forrest, 2008; LaGrange & Silverman, 1999). Therefore, having proper opportunities coupled with an underlying racial bias or subconscious illusory correlation (or perhaps simply low self-control) is required for there to be invariance among racially disparate behaviors in policing.

On the other hand, these findings could be interpreted as truly a lack of behavioral

invariance, thereby resonating with Becker's (1957) work on discrimination in labor markets. The results show that officers may have different "tastes" for discrimination based on the behavior under consideration. And so those identified as at risk for disparate policing may differ by the type of behavior under consideration. Adopting a taste-based or statistical discrimination-based perspective would lead some to conclude that there are specific factors that may help explain why a subset of officers engaged in disparate arrests while other officers engaged in disparate stops. For example, it could be that officer preferences are socialized and learned within workgroups such as squads in a police beat, which in turn leads to the disparities observed in one kind of behavior but not the other. As described by Crank (1998), this could come from a cultural transmission of attitudes and beliefs about racial minority groups, which would help explain the observed influence of one's peers in shaping disparate behavior as well.

Unfortunately, a formal test of these two competing theoretical explanations for the lack of behavioral invariance observed in this study is not possible due to the small sample size of officers who had engaged in either disparate arrests, stops, or both. Understanding what these patterns of disparate behavior are attributed to, whether it be due to changes in opportunity structures (e.g., loopholes identified by some officers in standard operating procedures) or because of cultural learning processes that exist within specific workgroups, can shed key insights into potential reforms. If the reason that officers engage in one kind of disparate behavior but not the other is due to opportunity structures, then changes in standard operating procedures and officer supervision may be a viable remediation strategy; otherwise, additional officer training (i.e., implicit bias training) and or shifts in workplace culture may be required. Future research should consider replicating these analyses in other agencies to test the external validity of this result and to provide additional insight into its practical implications for police reform.

This leads to the next set of findings that shed light on some policy-relevant questions about how agencies may respond to racial inequalities in police behavior. Interestingly, the results show that officers who engaged in racially-disparate stop and/or arrest behavior share a similar gender composition to those officers who did not engage in such behavior. The only consistent difference between these two sets of officers is that "bad apples" were, on average, a year younger than those who were not based on their stop behavior, whereas officers were a year older among bad apples in their arrest behavior. While this finding may lack practical significance—not many agencies will be too concerned about a one-year gap in officer age and its relation to disparate behavior—it does align with findings on the relationship between officers' age and the number of complaints and allegations of misconduct against them (Kane & White, 2009; Wolfe & Piquero, 2011).

Results from this study also show that the difference in the proportion of stops and arrests involving Black citizens is inconsistent and negligible when comparing male and female officers who do not engage in racially disparate behavior. In other words, there is no reliable evidence that officers' gender consistently predicts racial equity in their stop and arrest behavior. In contrast, results from this study show (albeit inconsistently) that White officers conducted more racially equitable stops than their Black colleagues when restricting attention to those officers who do not engage in racially-disparate behavior. Collectively, these findings contrast an empirically robust line of research that shows both female officers and Black officers may be less likely to perpetuate racial disparities in police outcomes (Ba et al., 2021; Gonçalves & Mello, 2021; Hoekstra & Sloan, 2022). However, it is crucial to note that there was a noticeable under-representation of Black officers and female officers in the analyzable sample of data. What these results may be speaking to are thus a subset of officers in the agency who arrest and stop more Chicagoans than most other Black officers and female officers in CPD. As such, the empirical investigation into whether

officers' race and gender matter in police agencies is far from complete. The results presented here are a snapshot of CPD's reality and should be investigated further in light of the inconsistent results reported here.

In light of these inconsistent findings observed in a single agency, research should consider how diversification policies theoretically and practically reap their intended effects on racially-disparate police behavior. Scholars advocating for more female representation in policing have largely relied on the idea that female "soft skills" explain the racial equity of their police behavior. Yet, given the lack of empirical support in this study, it may be worthwhile to explore the causal mechanisms that would generate this racial equity in police behavior.

One potential avenue worth exploring is how racial and gender identities are connected to racial equity based on socialization and lived experiences. Does experiencing social marginalization instigate an implicit or explicit preoccupation towards generating more equitable police outcomes? Alternatively, this could induce a gendered and racialized general strain stemming from identity conflicts, socialization pressures, and perceived discrimination (Isom & Grosholz, 2019; Isom & Mikell, 2019). For example, Isom and Grosholz (2019) show that while Black residents perceived injustices by police officers in Chicago were significantly related to criminogenic coping mechanisms, there was no such relationship among White residents. They attributed this to a racialized general strain specific to Black citizens' identities and their experiences with discrimination, which led to maladaptive responses towards society more generally. However, it is difficult to attribute racial and gender identities to the findings observed in this study without qualitative data to contextualize them. Future research should consider case studies and qualitative research that can tie officers' identities to their behavior and explore why such behavior is (or is not) associated with those identities.

Given the lack of consistent evidence suggesting officer demographics predict racially disparate and racially equitable behaviors, it may be that racial inequalities in policing are driven by a more systemic process, or at least shaped by situational factors. Indeed, when modeling police beat assignments for officers who were found to engage in racially-disparate stop and arrest behavior over four years, the results show that their behavior can be associated with their working environments. When these outliers work in police beats with higher concentrations of Black residents, they are far more likely to disproportionately stop and arrest Black civilians relative to their peers working the same beat on the same day that they are.

The results are encouraging for Clarke's (1995) theoretical propositions of situational crime prevention and proponents of environmental criminology and routine activity theory. In drawing from rational choice-based theories of organizational misconduct and discrimination in labor markets (Arrow, 1963; Becker, 1968; Greve, Palmer, & Pozner, 2010), I find that police officers' behavior is malleable based on opportunity contexts. Higher opportunity contexts are typically associated with a higher probability of an officer engaging in racially-disparate behavior. This also corresponds with Gonçalves and Mello (2021), who similarly found that Florida highway patrol officers were less likely to discriminate in their traffic ticketing practices when patrolling jurisdictions with fewer minority drivers on the roadways. More generally, this supports the broader criminological literature which has shown that criminal offending can be reduced by limiting access to desirable victims (for review, see Guerette & Bowers, 2009).

It is important to note that these findings are also met with some conflicting evidence. Officers were more likely to stop fewer Black civilians than their peers even when they had more opportunities to stop them. Racially-disparate behavior may therefore be sensitive to environmental contexts, but the observed patterns were inconsistent and, at times, counterintuitive.

What is needed moving forward is a deeper assessment of CPD officer behavior, perhaps through social systematic observations in body-worn cameras (e.g., McCluskey et al., 2018; Todak & James, 2018), to better understand the nature of this behavior and how they may develop perceptions of suspiciousness in different environments. Moreover, it might be worth exploring how much time officers spend actively on patrol when conducting disparity analyses. While some research shows that officers' decision-making may not be entirely motivated by making arrests towards the end of their shift to dip into overtime pay (Chalfin & Gonçalves, 2021), conducting arrests in and of themselves may reduce opportunities to engage in disparate behavior. When data are available, an important factor worth considering is thus how much time officers spend actively patrolling compared to filling out reports, making arrests, and other non-patrol-related activities.

Another finding worth noting here is the importance of peers in shaping officers' racially-disparate behavior. The results show that when an officer who was deemed a "bad apple" works in a beat where their fellow officers are also "bad apples" the probability of that officer engaging in racially-disparate arrests increases significantly. In contrast, the probability decreases for stopping Black civilians relative to their peers when working in police beats with fellow bad apple officers. These findings have been described by Warr (2002) in his discussion of the etiology of juvenile delinquency and peer relations: no longer is the question of *whether* peers matter but the question that remains is *how*. While this study shows that peer relations shape officers' propensity to engage in racially-disparate behavior, these findings corroborate a growing body of research that shows peer networks shape how officers learn about, perpetrate, and transmit misconduct in the Chicago Police Department (Ouellet et al., 2019). Importantly, the relationship between deviant peers and disparities was present but inconsistent. What is needed to move forward is more research on *how* peers matter in shaping officers' disparate behavior and when given that there is

some preliminary evidence to suggest that they can under certain circumstances and for certain behaviors.

Beyond the practical and theoretical findings discussed above, several methodological findings warrant further discussion. One of the challenges to studying racial differences in police behavior at a micro-level is that analysts must account for contextual factors if they seek to understand why some officers stop and arrest more Black civilians than others. The more robust approaches to this methodological challenge involve creating comparable situations in which officer behavior can be analyzed by finding unique ways to restrict variation in their data (Ba et al., 2021; Grogger & Ridgeway, 2006).

Isolating variation in data, however, is both a strength and a weakness. Data restrictions require analysts to make stringent assumptions about the importance of what information should be considered (and what should not) in a racial disparity analysis. This becomes especially important when considering how to make an officer-to-officer comparison when officers may conduct their daily work in vastly different working environments. The question at hand is thus how best can we create an apples-to-apples setting from which we can reasonably determine who engages in a disproportionately higher (or lower) rate of stops or arrests involving Black civilians relative to other officers?

In one recent and prominent study of racial disparities at a micro-level, Ba and colleagues (2021) place great importance on the month, day, beat, and shift assignment that each officer works—arguing that accounting for these factors creates a comparable setting from which officers' behavior can be analyzed. Yet, there still exist other factors that could influence whether some officers are predisposed to stopping Black civilians within these variables, such as in what neighborhood contexts they predominantly work within their police beats, what hour of the day

they are patrolling on a given shift, and what might be the reason they stop someone in the first place. Failing to account for these situational micro-temporal and spatial factors may mislead some into believing an officer is an outlier when in fact their behavior may be due to these objective factors.

In this study, I used a data-driven approach developed by Ridgeway and MacDonald (2009) to generate officer-to-officer comparisons in a racial disparity analysis. I argue that a data-driven approach may produce more useful models than a data-restrictive approach. This methodological approach is perhaps best motivated by Box and Draper's (1987) comments on model adequacy, with the important question being how wrong a model must be to not be useful. I argue that the most useful models in racial disparity analyses will draw on the most information readily available to reflect the lived reality of an officer in question. I rely on computational algorithms to determine how best to make the most of what information is available, rather than imposing assumptions on what information should be used.

I argue that imposing data restrictions can create useful models as well, but only if such restrictions are universally applicable to each officer's lived reality. In this way, Ba and colleagues' (2021) approach to generating officer-to-officer comparisons can be considered as a reduced form of the approach deployed in this study, where data restrictions can conform to analysts' expectations when the information corresponds to the reality of an officer in question. However, when the gradient boosting algorithm finds a better way to model an officer's reality, then my model will yield more robust inferences because it is a data-driven approach, not a data-restrictive one to generate comparisons. As recommended by Neil and Winship (2019) and demonstrated in this study, using matching models (and gradient boosting capabilities) can have the most potential to help agencies identify which officers engage in disparate behavior when the data permit—more

on this later.

Another important methodological contribution demonstrated in this study is the importance of accounting for the presumed risk of false positive identification in a racial disparity analysis. One of the challenges to conducting officer-to-officer comparisons in an internal benchmark analysis is that for every officer analyzed, the risk of a false-positive increase (see Equation 6). There are many approaches to adjusting for this multiple-hypothesis testing bias; however, this study demonstrated how controlling for the local false discovery rate offers analysts a unique opportunity to incorporate additional assumptions about the acceptable risk of falsely identifying an officer as an outlier relative to the risk of failing to identify an officer who is an outlier. As shown at the beginning of this study, when an agency places a greater concern for the risk of falsely identifying an officer as an outlier, the number of officers presumed to have engaged in disparate behavior shrinks dramatically.

If agencies are contemplating an internal benchmarking approach for their racial disparity analysis, they will need to come to terms with what they reasonably define as an acceptable level of risk for falsely identifying an officer as an outlier when they may not be one. When such benchmarking approaches are integrated into a proactive performance monitoring system, their thresholds can be modified to officers' behavior. For example, more potentially harmful behaviors (e.g., uses of force) can have more inclusive thresholds than less harmful ones (e.g., pedestrian stops). Here, the agency may be willing to accept higher chances of false positive identification given the behavior in question is more immediately consequential. Moreover, agencies under stricter public scrutiny may desire more inclusive thresholds than those who are more concerned with upsetting their organizational culture and climate. This offers a more flexible and transparent process by which agencies can assess potential racial disparities in their officers' behavior, while

also injecting theoretically relevant considerations that may substantially change their overall findings.

However, it bears noting that shifting the false discovery rate embodies a tradeoff between the potential gain/loss in police legitimacy externally and organizational justice internally. On one hand, implementing a more inclusive threshold will perhaps reduce the chances of bad apples skirting oversight and accountability measures, thereby restoring potential lost legitimacy in the eyes of the public when such behaviors impact them. On the other hand, this will likely socially martyrize officers and diminish their trust in the agency by being more willing to label them as potential contributors to a greater problem. Accordingly, agencies need to weigh the value of diminishing trust in the agency and gaining legitimacy in the eyes of the public when deciding on a false discovery rate – perhaps through algorithmic solutions such as linear programming to best determine the solution that meets their needs (Wheeler, 2020).

Study Limitations and Avenues for Future Research

This study is not without its limitations, which presents unique opportunities for future research to build upon. For starters, academics debate over whether discriminatory policing is the product of a "few bad apples." Some contend that the whole agency is corrupt, which would imply that making officer-to-officer comparisons for a racial disparity analysis is doomed from the start (Walsh, 2021). This is because the analysis would either generate no outliers (those officers whose behavior deviates statistically from others) or only identify the "worst of the worst" officers in an already corrupt system.

Results from this study present a partial rebuttal: some officers deviate from others in the racial composition of their stops with varying degrees of certainty as to how many. This suggests that there is no complete homogeneity of discriminatory behavior in this sample. Officers in this

sample exhibited considerable variation in their stop and arrest behavior, and that is reflected in the racial composition of their stops and arrests within and between the beats that they regularly work in. However, because the analytic procedure assumes that the majority of the sample does not engage in disparate behavior, either the findings suggest that there is truly a small fraction of officers as having engaged in disparate behavior, or that there is still a chance that the analyses conducted in this study identify only the "worst of the worst." This is a natural limitation to the internal benchmarking strategy as a means for identifying disparities.

Nevertheless, proponents of systemic police reform would contend that remediating the "worst of the worst" will do little to reduce racial discrimination in policing. In support of this stance, Kaplan and Chalfin (2021) show there is very little utility in reducing use-of-force complaints if the Chicago Police Department had replaced its top 10% of officers who generated the most complaints in the agency. The results of this study present a unique opportunity for further inquiry. Officers' racially-disparate behavior was not shaped by their personal factors, which would support the goals for more systemic reform efforts. Moreover, disparate behavior can be shaped by environmental contexts and situational circumstances, though this finding is not nearly as consistent nor intuitive as one might expect. However, the results are (at best) a mixed success for systemic reformists because there was no way of knowing whether other reform strategies such as implicit bias training could remediate those who were found to be perpetuating racial disparities. Broadly speaking, further evidence is needed to determine whether a power-few framework or a systemic reform framework is best suited for evidence-based police reform and racial disparities.

These findings and limitations shed light on some exciting opportunities for future research. What if officers' behavior is shaped by workgroup dynamics, cultural norms, and accountability mechanisms? Evidence suggests that deviant peers and opportunity contexts

influence officers' propensity to engage in disparate behavior. Can agency-based or community-based systemic controls inhibit officers' behavior, and is their behavior influenced by a more culturally normative process that manifests in their organizational environment or within their communities? How might unmeasurable factors such as citizen demeanor impact observed disparities? Answering these questions will require new methodologies, both quantitative and qualitative. This might include using social systematic observations of body-worn camera data, or conducting qualitative interviews with police officers during, before, and immediately after they conduct routine pedestrian or traffic stops. Digging deeper into the potential situational dynamics of police-citizen interactions, as well as the influence of squad and precinct processes through qualitative research can bolster disparity analyses by triangulating evidence, which in turn can generate more nuanced findings.

Apart from conceptual challenges with the identification of racial disparities, there are also some conceptual challenges to the remediation of those disparities explored in this study. Most notably, what are the immediate policy implications of officer diversification and reassignment? The findings provide indirect evidence against diversification policies as a potential solution to racially-disparate behavior given that officers' gender and race did not serve as potential risk factors. Some evidence suggests that there may be same-race leniency biases (Ba et al., 2021), but as this study has shown with the same data there may be inconsistent differences (or none at all) in how officers treat Black civilians regardless of their race and gender. This is, however, only observational evidence that does not bear the same policy implications nor empirical rigor that experimental evaluations of a diversification hiring initiative would. Thus, the verdict is still out as to whether such policies are effective at responding to racially-disparate police behavior.

Results highlighting the influence of police beat composition on officers' behavior are

more encouraging but met with equally important caution and insights for additional research. Officers who engage in disparate behavior over four years are less likely to engage in that behavior on any given day when they work in jurisdictions with fewer Black civilians living there. This does not provide evidence of whether re-assigning those officers to lower-risk jurisdictions can reduce their behavior. Without a program evaluation coupled with a robust randomized experiment, it is difficult to determine if officers' behavior may change because of feeling supervised and creating an experimental reactivity effect, or because they are re-assigned to lower-risk locations.

One potential avenue for future research is to assess the utility of performance monitoring systems, such as an early warning system, that can be used to allocate enhanced supervision among officers who may be at risk of engaging in disparate behavior (Carter et al., 2024). Deploying a randomized controlled trial of such an intervention system, informed by internal benchmark analyses deployed in this study, may go a long way towards understanding if and how officers' behavior may change to different accountability-enhancing or opportunity-reducing strategies to control disparate police behavior.

As with all studies, a final cautionary reminder should be made regarding the limitations of the data. As mentioned at the beginning of the results section, the original data obtained for this study and the analyzable datasets may represent a subset of officers in the CPD that do not perfectly embody the average officer. One potential explanation for the lack of consistency between data sources is that the data originally collected by Ba and colleagues (2021) are a subset of all data originally reported by CPD for their stops and arrests.

When looking at all arrests conducted between 2014 and 2016 in CPD (see Table 20), the number of arrests conducted per sworn officer was 20 arrests for the entire agency (244,644 arrests

÷ 11,965 officers ≈ 20 arrests per sworn officer). Meanwhile, there were only 19 arrests conducted per officer based on data from the general use dataset during the same time period, which is much less frequent than the average of almost 71 arrests per officer in the analyzable dataset of arrests. What this means more generally is that the analyzable sample of arresting officers may be more active than the average CPD officer. Unfortunately, it is not possible to determine whether the sample of stopping officers is less (or more) active than the average CPD officer due to a lack of publicly accessible department-level data on pedestrian and vehicle stops. Accordingly, it may be that the behavior observed in the general use data reflects more (or less) active officers than the average officer in CPD.

Table 20. Number of Arrests per Officer in Chicago Police Department (2014-2016) by Data Source

Dataset	Total Arrests	Total Officers	Arrests per Officer
Department-Wide Dataset	244,644	11,965	20
General Use Dataset	114,285	6,015	19
Analyzable Dataset	18,531	261	71

Notes: Department-wide dataset refers to original data collected directly from Chicago Open Data Portal. General Use Dataset refers to the pre-cleaned data used by Ba et al (2021). Analyzable Dataset refers to the dataset after all data cleaning for the internal benchmarking analysis. Total officers represent total sworn officers with arresting power according to 2016 LEMAS data. Arrests per officer are rounded to the nearest integer.

Another important limitation worth considering is how data-demanding the internal benchmarking analysis is. A minimum of 50 stops or arrests were needed to conduct the internal benchmark analysis for an officer in question. While many officers conducted more than 50 stops over the 4-year period, few officers conducted at least that many arrests during the same period. This hindered the statistical power of the analyses for the policy-relevant questions and perhaps the consistency of their findings as well. Moreover, without proper information regarding how much police discretion was used when making stops and arrests, it may be difficult to discern how

much the observed disparities are attributed to officers themselves.

For some, it might be far-fetched to assume that most CPD officers in the general use datasets did not conduct 50 arrests over a 4-year period, however, when looking at the number of arrests made per CPD officer using publicly available data from 2014-2016, Table 21 shows that each sworn officer made about 6-8 arrests each year.[35] A similar pattern emerges when looking at New York City Police Department (NYPD) officers between 2014-2016, in which sworn officers made about 9-10 arrests per year. Even if one were to assume that less than half of sworn officers in NYPD were considered regular patrol officers, there would still only be between 17 to 21 arrests per officer in any given year.[36] Similar results can be seen when examining the number of arrests per officer in the Los Angeles Police Department.[37]

Table 21. Number of Arrests per Officer by Agency (2014-2016)

City	Year	Total Arrests	Total Officers	Arrests per Officer
Chicago	2016	68,522	11,965	6
Chicago	2015	82,608	11,965	7
Chicago	2014	93,514	11,965	8
Los Angeles	2016	117,446	9,870	12
Los Angeles	2015	125,902	9,870	13
Los Angeles	2014	139,127	9,870	14
New York City	2016	313,473	36,634	9
New York City	2015	341,870	36,634	9
New York City	2014	380,600	36,634	10

Notes: Total officers represent total sworn officers with arresting power according to 2016 LEMAS data. Arrests per officer are rounded to the nearest integer and include both custody arrests and misdemeanor arrests.

What these results show collectively is that it might not be the arrest data that are

[35] The arrest data for Chicago come from an open data portal: https://data.cityofchicago.org/Public-Safety/Arrests/dpt3-jri9/data.
[36] Arrest data come from the City of New York open data portal: https://data.cityofnewyork.us/Public-Safety/NYPD-Arrests-Data-Historic-/8h9b-rp9u.
[37] Arrest data come from Los Angeles open data portal: https://data.lacity.org/Public-Safety/Arrest-Data-from-2010-to-2019/yru6-6re4/about_data.

unrepresentative of the agency, but that the internal benchmark analyses deployed in this study may be unrealistic for certain police behaviors in large agencies, and unrealistic altogether in smaller ones. Data analyzed in this study comes from one of the largest police departments in the country, and the frequency of arrests per officer is similar across other large police agencies. Accordingly, it is difficult to reconcile the utility of using the internal benchmarking analyses used here for smaller agencies if nothing else because most officers would simply not be applicable. If anything, the results from this study show that there is limited practical utility of the benchmarking procedure outside of stops.

The question becomes, what is the "best" approach to studying racial disparities in police agencies? Unfortunately, the answer is not absolute. For larger agencies, there are naturally more approaches to explore and more levels of analysis that they may consider. If a large agency deals with a fair amount of crime and traffic activity, then the internal benchmark stands a good chance of providing the agency with a precise estimate of who is most at risk for perpetuating racial disparities in pedestrian and traffic stops. This was the case for CPD, though it was not without its limitations.

While large agencies agency can deploy several kinds of racial disparity analyses conducted at higher levels of analysis such as a precinct, sector, or beat, smaller agencies will be restricted to these aggregate levels and specific types of analyses. This is because they have far fewer stops and arrests made by each officer, thus requiring them to analyze data at a larger level. As such, it would be ill-advised for small agencies to conduct an internal benchmark analysis as opposed to the Veil-of-Darkness method or one of the common non-experimental approaches.

In the end, there is no "silver bullet" racial disparity analysis for police agencies. Each has its own set of limitations. Agencies that can deploy multiple types of analyses will stand a better

chance at identifying potential racial disparities than those that cannot. Larger agencies stand the most to gain given the fact that they will have such large sample sizes in their stops, arrests, and use of force incidents. As shown in this study, they can also dig deeper into who may be perpetuating those disparities (if they exist). However, smaller agencies can conduct more than one kind of disparity analysis as well, such as coupling external benchmark analyses with a Veil-of-Darkness method. These will be restricted to an aggregate level such as assessing disparities across the entire agency. However, the best disparity analyses will depend on the data capacities of the agency. Accordingly, agencies should self-assess what type of analysis best suits the data infrastructure they have. It may also be worth considering new ways to collect data for racial disparities as well. Perhaps measuring disparities using novel body-worn camera footage and traffic dash cam footage may be one avenue towards solving the denominator problem that has plagued disparity analyses in policing since its inception.

These cautionary notes are shared not to discourage or invalidate the contributions of this study, but to remind the reader that these data are not perfect—as is the case in any study. In the end, this study shows how important it is to capture the lived realities of police officers' working environment if analysts want to understand who engaged in racially-disparate behavior. Rather than stopping there, this study further sheds light on some policy-relevant information that can help guide police reform evaluations that target the immediate objects of interest—the power few. I hope that such findings encourage scholars to continue their research on the study of racial disparities in police behavior and recognize the challenges to identification, and prospects of policy remediation, with the ultimate goal of finding ways to elevate the overall quality and fairness of police practices and procedures.

BIBLIOGRAPHY

Abraham, Y. (2023, March 22). Two bad apples? Let's look at the whole barrel. *Boston Globe*. https://www.bostonglobe.com/2023/03/22/metro/two-bad-apples-lets-look-whole-barrel/

Adams, I. T. (2021). *Modeling Officer Perceptions of Body-worn Cameras: A National Survey* [Preprint]. Thesis Commons. https://doi.org/10.31237/osf.io/fnxbg

Adams, I. T., McCrain, J., Schiff, D. S., Schiff, K. J., & Mourtgos, S. M. (2022). *Public Pressure or Peer Influence: What Shapes Police Executives' Views on Civilian Oversight?* [Preprint]. SocArXiv. https://doi.org/10.31235/osf.io/mdu96

Ali, M. U., & Pirog, M. (2019). Social Accountability and Institutional Change: The Case of Civilian Oversight of Police. *Public Administration Review*, *79*(3), 411–426. https://doi.org/10.1111/puar.13055

Allport, G. W. (1954). *The nature of prejudice*. Addison-Wesley.

Alpert, G. P., Smith, M. R., & Dunham, R. G. (2004). Toward a better benchmark: Assessing the utility of not-at-fault traffic crash data in racial profiling research. *Justice Research and Policy*, *6*(1), 43–69.

Ang, D. (2020). The Effects of Police Violence on Inner-City Students*. *The Quarterly Journal of Economics*, *136*(1), 115–168. https://doi.org/10.1093/qje/qjaa027

Anwar, S., & Fang, H. (2006). An alternative test of racial prejudice in motor vehicle searches: Theory and evidence. *American Economic Review*, *96*(1), 127–151.

Ariel, B., Sutherland, A., Henstock, D., Young, J., Drover, P., Sykes, J., Megicks, S., & Henderson, R. (2017). "Contagious Accountability": A Global Multisite Randomized Controlled Trial on the Effect of Police Body-Worn Cameras on Civilians' Complaints Against the Police. *Criminal Justice and Behavior*, *44*(2), 293–316. https://doi.org/10.1177/0093854816668218

Aronie, J., & Lopez, C. E. (2017). Keeping Each Other Safe: An Assessment of The Use of Peer Intervention Programs to Prevent Police Officer Mistakes and Misconduct, Using New Orleans' EPIC Program As A Potential National Model. *Police Quarterly*, *20*(3), 295–321. https://doi.org/10.1177/1098611117710443

Arrow, K. J. (1963). Uncertainty and the Welfare Economics of Medical Care. *The American Economic Review*, *53*(5), 941–973.

Ashforth, B. E., & Humphrey, R. H. (1997). The Ubiquity and Potency of Labeling in Organizations. *Organization Science*, *8*(1), 43–58. https://doi.org/10.1287/orsc.8.1.43

Atherley, L. T., & Hickman, M. J. (2013). Officer Decertification and the National Decertification Index. *Police Quarterly*, *16*(4), 420–437. https://doi.org/10.1177/1098611113489889

Ba, B. A., Knox, D., Mummolo, J., & Rivera, R. (2021). The role of officer race and gender in police-citizen interactions in Chicago. *Science*, *371*(6530), 696–702.

Becker, G. S. (1957). *The economics of discrimination*. University of Chicago press.

Becker, G. S. (1968). Crime and punishment: An economic approach. *Journal of political economy*, *76*(2), 169-217.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, *57*(1), 289-300.

Benton, M. (2020). Representation Is Not Enough: Symbolic Representation and Perceptions of the Police. *Administration & Society*, *52*(5), 794–822. https://doi.org/10.1177/0095399720905368

Bittner, E. (1967). The Police on Skid-Row: A Study of Peace Keeping. *American Sociological Review*, *32*(5), 699–715.

Black, D. J. (1973). The mobilization of law. *The Journal of Legal Studies*, *2*(1), 125–149.

Blalock Jr, H. M. (1967). Status inconsistency, social mobility, status integration and structural effects. *American Sociological Review*, 790–801.

Blumer, H. (1986). *Symbolic interactionism: Perspective and method*. Univ of California Press.

Blumstein, A. (1993). Making rationality relevant. The American society of criminology 1992 presidential address. *Criminology*, *31*(1), 1–16.

Blumstein, A., & Beck, A. J. (1999). Population growth in US prisons, 1980-1996. *Crime and Justice*, *26*, 17–61.

Bolger, P. C. (2015). Just Following Orders: A Meta-Analysis of the Correlates of American Police Officer Use of Force Decisions. *American Journal of Criminal Justice*, *40*(3), 466–492. https://doi.org/10.1007/s12103-014-9278-y

Bor, J., Venkataramani, A. S., Williams, D. R., & Tsai, A. C. (2018). Police killings and their spillover effects on the mental health of Black Americans: A population-based, quasi-experimental study. *The Lancet*, *392*(10144), 302–310.

Box, G. E., & Draper, N. R. (1987). *Empirical model-building and response surfaces*. John Wiley & Sons.

Braga, A. A., Brunson, R. K., & Drakulich, K. M. (2019). Race, place, and effective policing. *Annual Review of Sociology*, *45*(1), 535–555.

Braga, A. A., Papachristos, A. V., & Hureau, D. M. (2014). The effects of hot spots policing on crime: An updated systematic review and meta-analysis. *Justice Quarterly*, *31*(4), 633–663.

Brantingham, P. J., Valasik, M., & Mohler, G. O. (2018). Does predictive policing lead to biased arrests? Results from a randomized controlled trial. *Statistics and public policy*, 5(1), 1-6.

Brown, R. A., & Frank, J. (2006). Race and Officer Decision Making: Examining Differences in Arrest Outcomes between Black and White Officers. *Justice Quarterly, 23*(1), 96–126. https://doi.org/10.1080/07418820600552527

Brunson, R. K., & Gau, J. M. (2014). Race, Place, and Policing the Inner-City. In M. D. Reisig & R. J. Kane (Eds.), *The Oxford Handbook of Police and Policing*. Oxford University Press.

Brunson, R. K., & Wade, B. A. (2019). "Oh hell no, I don't talk to police" Insights on the lack of cooperation in police investigations of urban gun violence. *Criminology & Public Policy, 18*(3), 623–648.

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, *22*(1), 31–72.

Campeau, H. (2015). 'Police culture'at work: Making sense of police oversight. *British Journal of Criminology*, *55*(4), 669–687.

Carter T.M., Wolfe, S.E., Knode, J., & Henry, G. (2024). Attempting to Reduce Traffic Stop Racial Disparities: An Experimental Evaluation of an Internal Dashboard Intervention. *Criminology & Public Policy.* Forthcoming

CBS. (2019, August 7). *We asked 155 police departments about their racial bias training. Here's what they told us. - CBS News*. https://www.cbsnews.com/news/racial-bias-training-de-escalation-training-policing-in-america/

Chalfin, A., & Goncalves, F. (2021). The professional motivations of police officers. Unpublished manuscript.

Chalfin, A., & Kaplan, J. (2021). How many complaints against police officers can be abated by incapacitating a few "bad apples?" *Criminology & Public Policy*, *20*(2), 351–370.

Cho, S., Gonçalves, F., & Weisburst, E. (2021). Do Police Make Too Many Arrests? *Accessed On*, 06–10.

Christensen, T., & Connault, B. (2023). Counterfactual Sensitivity and Robustness. *Econometrica*, *91*(1), 263–298. https://doi.org/10.3982/ECTA17232

Christopher, W. (1991). *Report of the independent commission on the Los Angeles Police Department*. Diane Publishing.

Clarke, R. V. (1995). Situational crime prevention. *Crime and justice*, *19*, 91-150.

Clarke, S. (2009). Arrested Oversight: A Comparative Analysis and Case Study of How Civilian Oversight of the Police Should Function and How it Fails. *Columbia Journal of Law and*

Social Problems, *43*(1), 1–50.

Cobbina-Dungy, J. (2019). *Hands up, don't shoot: Why the protests in Ferguson and Baltimore matter, and how they changed America*. New York University Press.

Cobbina-Dungy, J. E., & Jones-Brown, D. (2023). Too much policing: Why calls are made to defund the police. *Punishment & Society*, *25*(1), 3–20.

Correll, J., & Keesee, T. (2009). Racial Bias in the Decision to Shoot? *The Police Chief*, *76*(5), 54–58.

Correll, J., Park, B., Judd, C. M., & Wittenbrink, B. (2002). The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals. *Journal of Personality and Social Psychology*, *83*(6), 1314–1329. https://doi.org/10.1037/0022-3514.83.6.1314

Correll, J., Park, B., Judd, C. M., Wittenbrink, B., Sadler, M. S., & Keesee, T. (2007). Across the thin blue line: Police officers and racial bias in the decision to shoot. *Journal of Personality and Social Psychology*, *92*(6), 1006–1023. https://doi.org/10.1037/0022-3514.92.6.1006

Council on Criminal Justice. (2021a). *Civilian Oversight* (Task Force on Policing: Policy Assessment, pp. 1–7). Council on Criminal Justice. https://assets.foleon.com/eu-west-2/uploads-7e3kk3/41697/civilian_oversight.2690411fd370.pdf

Council on Criminal Justice. (2021b). *Decertification* (Task Force on Policing: Policy Assessment, pp. 1–7). Council on Criminal Justice. https://assets.foleon.com/eu-west-2/uploads-7e3kk3/41697/decertification.d1229f8ea972.pdf

Council on Criminal Justice. (2021c). *Implicit Bias Training* (Task Force on Policing: Policy Assessment, pp. 1–6). Council on Criminal Justice. https://assets.foleon.com/eu-west-2/uploads-7e3kk3/41697/implicit_bias.524b7c301e55.pdf

Crank, J. P. (1998). *Understanding police culture*. Routledge.

Crank, J. P., & Langworthy, R. (1992). Institutional perspective on policing. *J. Crim. L. & Criminology*, *83*, 338.

Cullen, F. T., Link, B. G., Travis, L. F., & Lemming, T. (1983). Paradox in policing: A note on perceptions of danger. *Journal of Police Science & Administration*.

Cunningham, S. (2021). *Causal Inference: The Mixtape*. Yale University Press.

Dawson, A. J., Blount-Hill, K.-L., & Hodge Ii, G. (2022). Officer-involved deaths and the duty to intervene: Assessing the impact of DTI policy in New York City, 2000–2019. *Policing: An International Journal*, *45*(4), 662–675. https://doi.org/10.1108/PIJPSM-08-2021-0119

De Angelis, J., Rosenthal, R., & Buchner, B. (2016). *Civilian Oversight of Law Enforcement: A Review of the Strengths and Weaknesses of Various Models* (No. 250265; pp. 1–18). Office of Justice Programs.

DeAngelo, G., & Hansen, B. (2014). Life and death in the fast lane: Police enforcement and traffic fatalities. *American Economic Journal: Economic Policy*, *6*(2), 231–257.

Efron, B. (2004). The Estimation of Prediction Error: Covariance Penalties and Cross-Validation. *Journal of the American Statistical Association*, *99*(467), 619–632. https://doi.org/10.1198/016214504000000692

Eitle, D., D'Alessio, S. J., & Stolzenberg, L. (2002). Racial threat and social control: A test of the political, economic, and threat of Black crime hypotheses. *Social Forces*, *81*(2), 557–576.

Engel, R. S. (2008). A critique of the "outcome test" in racial profiling research. *Justice Quarterly*, *25*(1), 1–36.

Engel, R. S., McManus, H. D., & Herold, T. D. (2020). Does de-escalation training work?: A systematic review and call for evidence in police use-of-force reform. *Criminology & Public Policy*, *19*(3), 721–759. https://doi.org/10.1111/1745-9133.12467

Fagan, J., & Davies, G. (2000). Street stops and broken windows: Terry, race, and disorder in New York City. *Fordham Urb. LJ*, *28*, 457.

Felson, M., & Boba, R. L. (2010). *Crime and everyday life*. Sage.

Fliss, M. D., Baumgartner, F., Delamater, P., Marshall, S., Poole, C., & Robinson, W. (2020). Re-prioritizing traffic stops to reduce motor vehicle crash outcomes and racial disparities. *Injury Epidemiology*, *7*(1), 1–15.

French, D. (2023, February 5). 'Bad Apples' or Systemic Issues? *The New York Times*. https://www.nytimes.com/2023/02/05/opinion/memphis-police-academia-partisanship.html

Fryer, R. (2019). An Empirical Analysis of Racial Differences in Police Use of Force. *Journal of Political Economy*.

Gardner, A. M., & Scott, K. M. (2022). *Census of State and Local Law Enforcement Agencies, 2018 – Statistical Tables* (NCJ 302187; pp. 1–28). Bureau of Justice Statistics.

Gau, J. M., & Brunson, R. K. (2010). Procedural justice and order maintenance policing: A study of inner-city young men's perceptions of police legitimacy. *Justice Quarterly*, *27*(2), 255–279.

Gaub, J. E. (2020). Understanding Police Misconduct Correlates: Does Gender Matter in Predicting Career-Ending Misconduct? *Women & Criminal Justice*, *30*(4), 264–289. https://doi.org/10.1080/08974454.2019.1605561

Geller, A., Fagan, J., Tyler, T., & Link, B. G. (2014). Aggressive Policing and the Mental Health of Young Urban Men. *American Journal of Public Health*, *104*(12), 2321–2327. https://doi.org/10.2105/AJPH.2014.302046

Gelman, A., Fagan, J., & Kiss, A. (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.

Goel, S., Rao, J. M., & Shroff, R. (2016). Precinct or Prejudice? Understanding Racial Disparities in New York City's Stop-And-Frisk Policy. *The Annals of Applied Statistics*, *10*(1), 365–394.

Goeman, J. J., & Solari, A. (2014). Multiple hypothesis testing in genomics. *Statistics in Medicine*, *33*(11), 1946–1978. https://doi.org/10.1002/sim.6082

Goff, P. A. (2016). Identity traps: How to think about race & policing. *Behavioral Science & Policy*, *2*(2), 10–22.

Goldman, R., & Puro, S. (1987). Decertification of Police: An Alternative to Traditional Remedies for Police Misconduct. *Hastings Constitutional Law Quarterly*, *15*(45), 46–80.

Goldsmith, A. J. (2010). Policing's new visibility. *The British journal of criminology*, 50(5), 914-934.

Gonçalves, F., & Mello, S. (2021). A few bad apples? Racial bias in policing. *American Economic Review*, *111*(5), 1406–1441.

Gottfredson, M. R., & Hirschi, T. (1990). *A general theory of crime. Stanford University Press.*

Greve, H. R., Palmer, D., & Pozner, J. E. (2010). Organizations gone wild: The causes, processes, and consequences of organizational misconduct. *The Academy of Management Annals*, *4*(1), 53-107.

Grogger, J., & Ridgeway, G. (2006). Testing for Racial Profiling in Traffic Stops From Behind a Veil of Darkness. *Journal of the American Statistical Association*, *101*(475), 878–887. https://doi.org/10.1198/016214506000000168

Hadden, S. E. (2003). *Slave patrols: Law and violence in Virginia and the Carolinas*. Harvard University Press.

Hagan, J. (2012). Who are the criminals? In *Who Are the Criminals?* Princeton University Press.

Hamilton, D. L., & Gifford, R. K. (1976). Illusory correlation in interpersonal perception: A cognitive basis of stereotypic judgments. *Journal of Experimental Social Psychology*, *12*(4), 392-407.

Hay, C., & Forrest, W. (2008). Self-control theory and the concept of opportunity: The case for a more systematic union. *Criminology, 46(4), 1039-1072.*

Herbert, S. (1998). Police subculture reconsidered. *Criminology*, *36*(2), 343–370.

Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment

effects using the estimated propensity score. *Econometrica*, *71*(4), 1161–1189.

Hoekstra, M., & Sloan, C. (2022). Does Race Matter for Police Use of Force? Evidence from 911 Calls. *American Economic Review*, *112*(3), 827–860. https://doi.org/10.1257/aer.20201292

Holmes, M. D. (2000). Minority threat and police brutality: Determinants of civil rights criminal complaints in US municipalities. *Criminology*, *38*(2), 343–368.

Hope, K. R. (2020). Civilian oversight of the police: The case of Kenya. *The Police Journal: Theory, Practice and Principles*, *93*(3), 202–228. https://doi.org/10.1177/0032258X19860727

Huntington-Klein, N. (2021). *The Effect: An Introduction to Research Design and Causality*. CRC Press.

Isom Scott, D. A., & Grosholz, J. M. (2019). Unpacking the racial disparity in crime from a racialized general strain theory perspective. *Deviant Behavior, 40(12), 1445-1463.*

Isom Scott, D. A., & Mikell, T. (2019). 'Gender'and general strain theory: investigating the impact of gender socialization on young women's criminal outcomes. *Journal of Crime and Justice, 42(4), 393-413.*

Jacobs, D., & O'Brien, R. M. (1998). The determinants of deadly force: A structural analysis of police violence. *American Journal of Sociology*, *103*(4), 837–862.

James, L. (2018). The Stability of Implicit Racial Bias in Police Officers. *Police Quarterly*, *21*(1), 30–52. https://doi.org/10.1177/1098611117732974

James, L., James, S. M., & Vila, B. J. (2016). The Reverse Racism Effect: Are Cops More Hesitant to Shoot Black Than White Suspects? *Criminology & Public Policy*, *15*(2), 457–479. https://doi.org/10.1111/1745-9133.12187

James, L., Klinger, D., & Vila, B. (2014). Racial and ethnic bias in decisions to shoot seen through a stronger lens: Experimental results from high-fidelity laboratory simulations. *Journal of Experimental Criminology*, *10*(3), 323–340. https://doi.org/10.1007/s11292-014-9204-9

Johnson, D., Cesario, J., & Pleskac, T. (2018). How prior information and police experience impact decisions to shoot. *Journal of Personality and Social Psychology*, *115*(4), 601–623.

Johnson, T., & Johnson, N. (2023, February 3). If We Want to Reduce Deaths at Hands of Police, We Need to Reduce Traffic Stops. *TIME*. https://time.com/6252760/reducing-fatal-police-encounters-traffic-stops/

Jones-Brown, D. (2007). Forever the symbolic assailant: The more things change, the more they remain the same. *Criminology & Pub. Pol'y*, *6*, 103.

Jones-Brown, D., Dawson, A., Blount-Hill, K. L., Fuller, K. M., Oder, P., & Fradella, H. F. (2021). Am I my brother's keeper? Can duty to intervene policies save lives and reduce the need for special prosecutors in officer-involved homicide cases? *Criminal Justice Studies*, *34*(3), 306–351. https://doi.org/10.1080/1478601X.2021.1964694

Kalven v. City of Chicago, 09-CH-51396 (Illinois State Trial Court August 12, 2014).

Kane, R. J., & White, M. D. (2009). A study of career-ending misconduct among New York City police officers. *Criminology & Public Policy*, *8*(4), 737–769.

Keiser, L. R., Wilkins, V. M., Meier, K. J., & Holland, C. A. (2002). Lipstick and logarithms: Gender, institutional context, and representative bureaucracy. *American Political Science Review*, *96*(3), 553–564.

Kent, S. L., & Jacobs, D. (2005). Minority threat and police strength from 1980 to 2000: A fixed-effects analysis of nonlinear and interactive effects in large US cities. *Criminology*, *43*(3), 731–760.

Kirk, D. S., & Papachristos, A. V. (2011). Cultural mechanisms and the persistence of neighborhood violence. *American Journal of Sociology*, *116*(4), 1190–1233.

Klinger, D. (2004). *Into the Kill Zone: A Cop's Eye View of Deadly Force*. John Wiley & Sons.

Klinger, D. (1997). Negotiating order in patrol work: An ecological theory of police response to deviance. *Criminology*, *35*(2), 277–306.

Knode, J., Wolfe, S.E., & Carter, T.M. (2024). Shedding light on the Veil of Darkness: How to use the veil of darkness to examine racial disparity. *Criminology*. Forthcoming

Knowles, J., Persico, N., & Todd, P. (2001). Racial bias in motor vehicle searches: Theory and evidence. *Journal of Political Economy*, *109*(1), 203–229.

Korre, M., Farioli, A., Varvarigou, V., Sato, S., & Kales, S. N. (2014). A survey of stress levels and time spent across law enforcement duties: Police chief and officer agreement. *Policing: A Journal of Policy and Practice*, *8*(2), 109–122.

Kubrin, C. E., Squires, G., & Stewart, E. (2007). Neighborhoods, race, and recidivism: The Community reoffending nexus and its implications for African Americans. *SAGE Race Relations Abstracts*, *32*, 7–37.

Lai, C. K., & Lisnek, J. A. (2023). The Impact of Implicit-Bias-Oriented Diversity Training on Police Officers' Beliefs, Motivations, and Actions. *Psychological Science*, *34*(4), 424–434.

LaGrange, T. C., & Silverman, R. A. (1999). Low self-control and opportunity: Testing the general theory of crime as an explanation for gender differences in delinquency. *Criminology, 37(1), 41-72.*

Lange, J. E., Johnson, M. B., & Voas, R. B. (2005). Testing the racial profiling hypothesis for

seemingly disparate traffic stops on the New Jersey Turnpike. *Justice Quarterly*, *22*(2), 193–223.

Lavin, N. (2023, February 10). Oscar Perez named Providence police chief. *Providence Business News*. https://pbn.com/oscar-perez-named-providence-police-chief/

Legewie, J., & Fagan, J. (2019). Aggressive Policing and the Educational Performance of Minority Youth. *American Sociological Review*.

Lipsky, M. (1980). *Street-Level Bureaucracy, 30th Anniversary Edition: Dilemmas of the Individual in Public Service*. Russell Sage Foundation.

Liska, A. E. (1992). *Social threat and social control*. Suny Press.

Lum, C. (2009). Translating police research into practice. *Ideas in American Policing*, *11*(8), 1–15.

Lum, C. (2021). Perspectives on Policing: Cynthia Lum. *Annual Review of Criminology*, *4*, 19–25.

Lum, C., Koper, C. S., Wilson, D. B., Stoltz, M., Goodier, M., Eggins, E., Higginson, A., & Mazerolle, L. (2020). Body-worn cameras' effects on police officers and civilian behavior: A systematic review. *Campbell Systematic Reviews*, *16*(3). https://doi.org/10.1002/cl2.1112

Lum, C. M., & Koper, C. S. (2017). *Evidence-based policing: Translating research into practice*. Oxford University Press Oxford.

MacDonald, J. M., Kaminski, R. J., Alpert, G. P., & Tennenbaum, A. N. (2001). The temporal relationship between police killings of civilians and criminal homicide: A refined version of the danger-perception theory. *Crime & Delinquency*, *47*(2), 155–172.

Makowsky, M. D., & Stratmann, T. (2011). More Tickets, Fewer Accidents: How Cash-Strapped Towns Make for Safer Roads. *The Journal of Law and Economics*, *54*(4), 863–888. https://doi.org/10.1086/659260

Manning, P. K. (2007). A dialectic of organisational and occupational culture. *Sociology of Crime, Law and Deviance*, *8*, 47–83.

Manski, C. F., & Nagin, D. S. (2017). Assessing benefits, costs, and disparate racial impacts of confrontational proactive policing. *Proceedings of the National Academy of Sciences*, *114*(35), 9308–9313.

McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity Score Estimation With Boosted Regression for Evaluating Causal Effects in Observational Studies. *Psychological Methods*, *9*(4), 403–425. https://doi.org/10.1037/1082-989X.9.4.403

McCarthy, B., Hagan, J., & Herda, D. (2020). Neighborhood climates of legal cynicism and complaints about abuse of police power †. *Criminology*, *58*(3), 510–536.

https://doi.org/10.1111/1745-9125.12246

McCarthy, D. J. (2013). Gendering 'Soft' policing: Multi-agency working, female cops, and the fluidities of police culture/s. *Policing and Society, 23*(2), 261–278. https://doi.org/10.1080/10439463.2012.703199

McCarthy, J. (2022, May 27). *Americans Remain Steadfast on Policing Reform Needs in 2022*. Gallup.Com. https://news.gallup.com/poll/393119/americans-remain-steadfast-policing-reform-needs-2022.aspx

McCluskey, J. D., Uchida, C. D., Solomon, S. E., Wooditch, A., Connor, C., & Revier, L. (2019). Assessing the effects of body-worn cameras on procedural justice in the Los Angeles Police Department. *Criminology, 57(2), 208-236.*

McCrary, J. (2007). The Effect of Court-Ordered Hiring Quotas on the Composition and Quality of Police. *American Economic Review*, *97*(1), 318–353. https://doi.org/10.1257/aer.97.1.318

McDowell, M. G., & Fernandez, L. A. (2018). 'Disband, disempower, and disarm': Amplifying the theory and practice of police abolition. *Critical Criminology*, *26*, 373–391.

McLean, K., Wolfe, S. E., Rojek, J., Alpert, G. P., & Smith, M. R. (2020). Randomized controlled trial of social interaction police training. *Criminology & Public Policy, 19*(3), 805–832. https://doi.org/10.1111/1745-9133.12506

Mead, G. H. (1934). *Mind, self, and society* (Vol. 111). Chicago: University of Chicago press.

Mears, D. P., Craig, M. O., Stewart, E. A., & Warren, P. Y. (2017). Thinking fast, not slow: How cognitive biases may contribute to racial disparities in the use of force in police-citizen encounters. *Journal of Criminal Justice, 53*, 12–24.

Meier, K. J., & Nicholson-Crotty, J. (2006). Gender, representative bureaucracy, and law enforcement: The case of sexual assault. *Public Administration Review*, *66*(6), 850–860.

Monkkonen, E. H. (2004). *Police in urban America, 1860-1920*. Cambridge University Press.

Muir, W. K. (1977). *Police: Streetcorner politicians*. University of Chicago Press.

Nam, Y., Wolfe, S. E., & Nix, J. (2022). Does Procedural Justice Reduce the Harmful Effects of Perceived Ineffectiveness on Police Legitimacy? *Journal of Research in Crime and Delinquency*, 002242782211216. https://doi.org/10.1177/00224278221121622

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, *7*. https://doi.org/10.3389/fnbot.2013.00021

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

National Research Council. (2004). *Measuring racial discrimination*. National Academies Press.

Neil, R., & Winship, C. (2019). Methodological challenges and opportunities in testing for racial discrimination in policing. *Annual Review of Criminology*, *2*, 73–98.

Nguyen, V., & Ridgeway, G. (2023). Judges on the Benchmark: Developing a Sentencing Feedback System. *Justice Quarterly*, 1-37.

Nix, J., Huff, J., Wolfe, S. E., Pyrooz, D. C., & Mourtgos, S. M. (2024). When police pull back: Neighborhood-level effects of de-policing on violent and property crime, a research note. *Criminology*.

Novak, K. J., & Chamlin, M. B. (2012). Racial threat, suspicion, and police behavior: The impact of race and place in traffic enforcement. *Crime & Delinquency*, *58*(2), 275–300.

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

Paoline III, E. A. (2003). Taking stock: Toward a richer understanding of police culture. *Journal of Criminal Justice*, *31*(3), 199–214.

Petrocelli, M., Piquero, A. R., & Smith, M. R. (2003). Conflict theory and racial profiling: An empirical analysis of police traffic stop data. *Journal of Criminal Justice*, *31*(1), 1–11.

Peyton, K. (2021). What can we learn about police attitudes from four decades of the General Social Survey? A comment on Roscigno and Preito-Hodge (2021). *Pre-print.*

Plant, E. A., & Peruche, B. M. (2005). The Consequences of Race for Police Officers' Responses to Criminal Suspects. *Psychological Science*, *16*(3), 180–183. https://doi.org/10.1111/j.0956-7976.2005.00800.x

President's Task Force on 21st Century Policing. (2015). *Final Report of the President's Task Force on 21st Century Policing* (pp. 1–117). Office of Community Oriented Policing Services.

Puntis, S., Perfect, D., Kirubarajan, A., Bolton, S., Davies, F., Hayes, A., Harriss, E., & Molodynski, A. (2018). A systematic review of co-responder models of police mental health 'street' triage. *BMC Psychiatry*, *18*(1), 256. https://doi.org/10.1186/s12888-018-1836-2

Reichel, P. L. (1988). Southern slave patrols as a transitional police type. *Am. J. Police*, *7*, 51.

Reisig, M. D. (2010). Community and problem-oriented policing. *Crime and justice*, *39*(1), 1-53.

Rengifo, A. F., & Fowler, K. (2016). Stop, question, and complain: Civilian grievances against the NYPD and the opacity of police stops across New York City precincts, 2007–2013. *Journal of Urban Health*, *93*, 32–41.

Reuland, M. (2010). Tailoring the police response to people with mental illness to community

characteristics in the USA. *Police Practice and Research*, *11*(4), 315–329. https://doi.org/10.1080/15614261003701723

Riccucci, N. M., Van Ryzin, G. G., & Jackson, K. (2018). Representative Bureaucracy, Race, and Policing: A Survey Experiment. *Journal of Public Administration Research and Theory*, *28*(4), 506–518. https://doi.org/10.1093/jopart/muy023

Riccucci, N. M., Van Ryzin, G. G., & Lavena, C. F. (2014). Representative Bureaucracy in Policing: Does It Increase Perceived Legitimacy? *Journal of Public Administration Research and Theory*, *24*(3), 537–551. https://doi.org/10.1093/jopart/muu006

Richardson, R., Schultz, J. M., & Crawford, K. (2019). Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. *NYUL Rev.*, 94, 15.

Ridgeway, G. (2006). Assessing the effect of race bias in post-traffic stop outcomes using propensity scores. *Journal of quantitative criminology*, 22, 1-29.

Ridgeway, G. (2018). Policing in the era of big data. *Annual Review of Criminology*, *1*, 401–419.

Ridgeway, G., & MacDonald, J. (2010). Methods for assessing racially biased policing. In S. K. Rice & M. D. White (Eds.), *Race, ethnicity, and policing: New and essential readings*. New York University Press.

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

Ridgeway, G., & MacDonald, J. M. (2014). A method for internal benchmarking of criminal justice system performance. *Crime & Delinquency*, *60*(1), 145-162.

Ridgeway, G., Moyer, R. A., & Bushway, S. D. (2020). Sentencing scorecards: Reducing racial disparities in prison sentences at their source. *Criminology & Public Policy*, *19*(4), 1113-1138.

Roach, K., Baumgartner, F. R., Christiani, L., Epp, D. A., & Shoub, K. (2022). At the intersection: Race, gender, and discretion in police traffic stop outcomes. *Journal of Race, Ethnicity, and Politics*, 7(2), 239-261.

Rojek, J., Rosenfeld, R., & Decker, S. (2012). Policing race: The racial stratification of searches in police traffic stops. *Criminology, 50(4), 993-1024.*

Roscigno, V. J., & Preito-Hodge, K. (2021). Racist cops, vested "blue" interests, or both? Evidence from four decades of the General Social Survey. *Socius*, *7*, 2378023120980913.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41-55.

Rowe, M. (2023). *Disassembling Police Culture*. Taylor & Francis.

Rozema, K., & Schanzenbach, M. (2019). Good cop, bad cop: Using civilian allegations to predict police misconduct. *American Economic Journal: Economic Policy*, *11*(2), 225–268.

Saad, G. (2022, May 19). *Concern About Race Relations Persists After Floyd's Death*. Gallup.Com. https://news.gallup.com/poll/392705/concern-race-relations-persists-floyd-death.aspx

Sadler, M. S., Correll, J., Park, B., & Judd, C. M. (2012). The World Is Not Black and White: Racial Bias in the Decision to Shoot in a Multiethnic Context: Multiethnic Racial Bias. *Journal of Social Issues*, *68*(2), 286–313. https://doi.org/10.1111/j.1540-4560.2012.01749.x

Sampson, R. J., & Lauritsen, J. L. (1997). Racial and ethnic disparities in crime and criminal justice in the United States. *Crime and Justice*, *21*, 311–374.

Sampson, R. J., Raudenbush, S. W., & Earls, F. (1997). Neighborhoods and violent crime: A multilevel study of collective efficacy. *Science*, *277*(5328), 918–924.

Saucier v. Katz, No. 99-1977 (United States Supreme Court 2001).

Schuck, A. M., & Rabe-Hemp, C. (2005). Women Police: The Use of Force by and Against female Officers. *Women & Criminal Justice*, *16*(4), 91–117. https://doi.org/10.1300/J012v16n04_05

Sewell, A. A., & Jefferson, K. A. (2016). Collateral Damage: The Health Effects of Invasive Police Encounters in New York City. *Journal of Urban Health*, *93*(S1), 42–67. https://doi.org/10.1007/s11524-015-0016-7

Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology*, *46*(1), 561–584.

Shear, M. D., Tankersley, J., & Kanno-Youngs, Z. (2023, February 8). 7 Takeaways From Biden's State of the Union Address. *The New York Times*. https://www.nytimes.com/2023/02/07/us/politics/biden-state-of-the-union-takeaways.html

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

Sherman, L. W. (2007). The power few: Experimental criminology and the reduction of harm: The 2006 Joan McCord Prize Lecture. *Journal of experimental criminology*, *3*, 299-321.

Sherman, L. W. (2013). The rise of evidence-based policing: Targeting, testing, and tracking. *Crime and Justice*, *42*(1), 377–451.

Shoub, K. (2023, March 20). Traffic stop data shows that police reform should focus on remaking the barrel, not just addressing "bad apples." *Justice and Domestic Affairs*. https://blogs.lse.ac.uk/usappblog/2023/03/20/traffic-stop-data-shows-that-police-reform-should-focus-on-remaking-the-barrel-not-just-addressing-bad-apples/

Sierra-Arévalo, M. (2021). American policing and the danger imperative. *Law & Society Review*, *55*(1), 70–103.

Sierra-Arévalo, M. (2021). Reward and "Real" Police Work. In B. Jones & E. Mendieta (Eds.), *The ethics of policing: New perspectives on law enforcement*. New York University Press.

Skolnick, J. (1966). Justice without Trial: Law Enforcement in Democratic Society. *Englewood Cliffs: McMillian*.

Smith, M. R., & Alpert, G. P. (2007). Explaining police bias: A theory of social conditioning and illusory correlation. *Criminal Justice and Behavior*, *34*(10), 1262–1283.

Smith, M. R., Makarios, M., & Alpert, G. P. (2006). Differential suspicion: Theory specification and gender effects in the traffic stop context. *Justice Quarterly*, *23*(02), 271-295.

Smith, M. R., Tillyer, R., Lloyd, C., & Petrocelli, M. (2021). Benchmarking disparities in police stops: A comparative application of 2nd and 3rd generation techniques. *Justice Quarterly*, *38*(3), 513–536.

Spohn, C., & Holleran, D. (2002). The effect of imprisonment on recidivism rates of felony offenders: A focus on drug offenders. *Criminology*, *40*(2), 329–358.

Stults, B. J., & Baumer, E. P. (2007). Racial context and police force size: Evaluating the empirical validity of the minority threat perspective. *American Journal of Sociology*, *113*(2), 507–546.

Taniguchi, T., Vovak, H., Cordner, G., Amendola, K., Yang, Y., Hoogesteyn, K., & Bartness, M. (2022). The Impact of Active Bystander Training on Officer Confidence and Ability to Address Ethical Challenges. *Policing: A Journal of Policy and Practice*, *16*(3), 508–522. https://doi.org/10.1093/police/paac034

Taylor, P. L. (2020). Dispatch Priming and the Police Decision to Use Deadly Force. *Police Quarterly*, *23*(3), 311–332. https://doi.org/10.1177/1098611119896653

Terrill, W., & Ingram, J. R. (2016). Civilian Complaints Against the Police: An Eight City Examination. *Police Quarterly*, *19*(2), 150–179. https://doi.org/10.1177/1098611115613320

Terrill, W., & Reisig, M. D. (2003). Neighborhood context and police use of force. *Journal of Research in Crime and Delinquency*, *40*(3), 291–321.

Thacher, D. (2022). Shrinking the police footprint. *Criminal Justice Ethics, 41(1), 62-85.*

Theobald, N. A., & Haider-Markel, D. P. (2008). Race, Bureaucracy, and Symbolic Representation: Interactions between Civilians and Police. *Journal of Public Administration Research and Theory*, *19*(2), 409–426. https://doi.org/10.1093/jopart/mun006

Theobald, N. A., & Haider-Markel, D. P. (2009). Race, bureaucracy, and symbolic representation: Interactions between civilians and police. *Journal of Public Administration*

Research and Theory, *19*(2), 409–426.

Thomas, W. I., & Thomas, D. S. (1928). *The child in America: Behavior problems and programs*. Knopf.

Todak, N., & James, L. (2018). A Systematic Social Observation Study of Police De-Escalation Tactics. *Police Quarterly*, *21*(4), 509–543. https://doi.org/10.1177/1098611118784007

Tomaskovic-Devey, D., Mason, M., & Zingraff, M. (2004). Looking for the driving while Black phenomena: Conceptualizing racial bias processes and their associated distributions. *Police Quarterly, 7*(1), 3–29.

Tomkins, S. S. (2008). *Affect imagery consciousness: the complete edition: two volumes*. Springer publishing company.

Tonry, M. (2011). *Punishing race: A continuing American dilemma*. Oxford University Press.

Trojanowicz, R. C. (1983). An evaluation of a neighborhood foot patrol program. *Journal of Police Science & Administration*.

Tyler, T. R., & Fagan, J. (2008). Legitimacy and cooperation: Why do people help the police fight crime in their communities. *Ohio St. J. Crim. L.*, *6*, 231.

Tyler, T. R., Jackson, J., & Mentovich, A. (2015). The consequences of being an object of suspicion: Potential pitfalls of proactive police contact. *Journal of Empirical Legal Studies*, *12*(4), 602–636.

United States Department of Justice Civil Rights Division. (2015). *The Ferguson Report: Department of Justice Investigation of the Ferguson Police Department*. New Press, The.

Van Maanen, J. (1978). The asshole. *Policing: A View from the Street*, 221–238.

Vito, A. G., Woodward Griffin, V., Vito, G. F., & Higgins, G. E. (2020). "Does daylight matter"? An examination of racial bias in traffic stops by police. *Policing: An International Journal*, *43*(4), 675–688. https://doi.org/10.1108/PIJPSM-04-2020-0055

Waddington, P. A. (1999). Police (canteen) sub-culture. An appreciation. *The British Journal of Criminology*, *39*(2), 287–309.

Waernbaum, I. (2012). Model misspecification and robustness in causal inference: comparing matching with doubly robust estimation. *Statistics in medicine*, 31(15), 1572-1581.

Walker, S. (2001). Searching for the denominator: Problems with police traffic stop data and an early warning system solution. *Justice Research and Policy*, *3*(1), 63–95.

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

Walsh, Colleen. 2021. “Solving Racial Disparities in Policing.” *Harvard Gazette*. Retrieved January 18, 2024 (https://news.harvard.edu/gazette/story/2021/02/solving-racial-disparities-in-policing/).

Warr, M. (2002). *Companions in crime: The social aspects of criminal conduct*. Cambridge University Press.

Webb, C., Linn, S., & Lebo, M. (2019). A Bounds Approach to Inference Using the Long Run Multiplier. *Political Analysis*, *27*(3), 281–301. https://doi.org/10.1017/pan.2019.3

Wehrman, M. M. (2010). Race, concentrated disadvantage, and recidivism: A test of interaction effects. *Journal of Criminal Justice*, *38*(4), 538–544.

Weisburd, D., & Green, L. (1995). Policing drug hot spots: The Jersey City drug market analysis experiment. *Justice Quarterly*, *12*(4), 711-735.

Weisburst, E. K. (2019). Patrolling Public Schools: The Impact of Funding for School Police on Student Discipline and Long-term Education Outcomes: Patrolling Public Schools. *Journal of Policy Analysis and Management*, *38*(2), 338–365. https://doi.org/10.1002/pam.22116

Weitzer, R. (2015). American policing under fire: Misconduct and reform. *Society*, *52*, 475–480.

West, J. (2018). Racial Bias in Police Investigations. *NBER*.

Wheeler, A. P. (2020). Allocating police resources while limiting racial inequality. *Justice quarterly*, 37(5), 842-868.

White, C., & Weisburd, D. (2018). A Co-Responder Model for Policing Mental Health Problems at Crime Hot Spots: Findings from a Pilot Project. *Policing: A Journal of Policy and Practice*, *12*(2), 194–209. https://doi.org/10.1093/police/pax010

White, D. R., Schafer, J., & Kyle, M. (2022). The impact of COVID-19 on police training academies. *Policing: An International Journal*, *45*(1), 9–22. https://doi.org/10.1108/PIJPSM-06-2021-0078

White, M. D., & Malm, A. (2020). *Cops, Cameras, and Crisis: The Potential and the Perils of Police Body-Worn Cameras*. NYU Press.

Wilson, W. J. (1987). *The truly disadvantaged: The inner city, the underclass, and public policy*. University of Chicago Press.

Witkin, N. (2016). The Police-Community Partnership: Civilian Oversight as an Evaluation Tool for Community Policing. *St. Mary’s Law Review on Race and Social Justice*, *18*(2).

Wojcicki, E. (2022). *Illinois survey finds crisis in police recruitment and retention*. Police1. https://www.police1.com/police-recruiting/articles/illinois-survey-finds-crisis-in-police-recruitment-and-retention-tiq5if6IbrHafPuM/

Wolfe, S. E., & Lawson, S. G. (2020). The organizational justice effect among criminal justice employees: A meta-analysis. *Criminology*, *58*(4), 619-644.

Wolfe, S. E., & Nix, J. (2017). Police Officers' Trust in Their Agency: Does Self-Legitimacy Protect Against Supervisor Procedural Injustice? *Criminal Justice and Behavior*, *44*(5), 717–732. https://doi.org/10.1177/0093854816671753

Wolfe, S. E., & Piquero, A. R. (2011). Organizational justice and police misconduct. *Criminal justice and behavior*, *38*(4), 332-353.

Wolfe, S., Rojek, J., McLean, K., & Alpert, G. (2020). Social Interaction Training to Reduce Police Use of Force. *The ANNALS of the American Academy of Political and Social Science*, *687*(1), 124–145. https://doi.org/10.1177/0002716219887366

Wolfgang, M. E., Figlio, R. M., & Sellin, T. (1972). *Delinquency in a birth cohort*. University of Chicago Press.

Wooldridge, J. M. (2015). *Introductory Econometrics: A Modern Approach*. Cengage Learning.

Worden, R. E., McLean, S. J., & Wheeler, A. P. (2012). Testing for Racial Profiling With the Veil-of-Darkness Method. *Police Quarterly*, *15*(1), 92–111. https://doi.org/10.1177/1098611111433027

Wrigley-Field, E. (2020). Life Years Lost to Police Encounters in the United States. *Socius: Sociological Research for a Dynamic World*, *6*, 237802312094871. https://doi.org/10.1177/2378023120948718