Race and the Ecology of Policing: Testing for Neighborhood-Level Discrimination in the NYPD's Program of Stop, Question, and Frisk

Roland Neil

Abstract: Academic and popular explanations of police behavior tend to emphasize the importance of individual and incident-level factors, especially whether officers target individuals of specific races and associated processes like implicit bias. However, in order to understand police behavior—including but not limited to racial disparities that result from it—it is essential to understand why officers behave so differently in different neighborhoods. I argue that the racial composition of neighborhoods is an important driver of police behavior—what I call ecological discrimination—and describe four specific mechanisms as to why. While prior research has examined racial composition in relation to police behavior, I discuss several serious challenges facing this body of work. Using data on nearly 750,000 weapon stops conducted by the NYPD between 2008 and 2012, I implement a novel matching approach to test whether ecological discrimination explains spatial variation in police behavior. Results indicate that ecological discrimination was an important driver of NYPD stop patterns. These results carry important implications for understanding police behavior and for the proper design of studies that relate racial composition to police behavior, and the matching approach employed offers a means of making stronger inferences as to how context affects individuals.

Recently, there has been intense focus and debate on why American police behave the way they do, especially as to why racial disparities in various forms of police contact exist. In this context, the New York Police Department's (NYPD) practice of Stop, Question, and Frisk (SQF) has emerged as arguably the single most well-known and contested case. Academics, journalists, courts, and even presidential candidates have all weighed in on how to explain SQF (e.g.

Borchetta, Charney, & Harris, 2018; Fagan, 2017a; *Floyd et al. v. City of New York*, 2013; Ridgeway, 2017). Mirroring debates around the causes of police behavior more generally, these explanations have tended to focus on the importance of individual and incident-level features. Most often, they have focused on the extent to which police officers' decisions to stop suspects are driven by racial bias as opposed to legitimate attempts to proactively combat crime (Gelman, Fagan, & Kiss, 2007; Goel, Rao, & Shroff, 2016; Ridgeway, 2007; Zimring, 2011).

However, because of this focus on individual and incident-level factors, existing research cannot account for many of the observed patterns in NYPD stops. Both the number of stops and the typical standards of suspicion that the police employ in deciding to make a stop vary massively depending on the neighborhood that they are in, and this is not solely a result of differences in the people that they encounter in different places. In order to understand why the NYPD stopped who they did, including racial disparities in their stop patterns, it is necessary to understand why they behave so differently in different places. In this article, I advance ecological discrimination—that police discriminate at the level of the neighborhood—as a key explanation of spatial variation in NYPD stop patterns. That is, the racial composition of an area can affect police behavior, and I highlight four specific mechanisms as to why this might be the case which pertain to how context shapes cognition, social disorganization, the use and policing of public spaces, and racial threat.

Many studies have examined racial composition in relation to police behavior (e.g. Carmichael & Kent, 2014; Fagan, Geller, Davies, & West, 2010; Jacobs & O'Brien, 1998; MacDonald & Braga, 2018). I show several challenges that cast doubt on the findings of such analyses. These studies are often not careful to disentangle an emergent racial composition effect from the effect of the race of individuals found in different places, nor do they tend to control for competing individual, incident, and ecological-level explanations. Moreover, these studies rely exclusively on regression in a context where it is unlikely to perform well. I implement a novel matching approach that offers better performance in this context and use it to guard against competing explanations so as to isolate a racial composition effect.

I find ecological discrimination to be an important explanation of SQF. In fact, the racial context of areas is at least as important as the race of individuals in explaining variation in the typical standards of suspicion employed when making stops. These findings constitute an important explanation of Stop, Question, and Frisk, including the racial disparities that resulted from it. This analysis highlights that neighborhood-level variation in police behavior is a central thing to be explained if we want to understand why police do what they do, not something to be controlled away. Additionally, it suggests the large body of research relating racial composition to police behavior must be far more careful in the research designs used, and it presents a novel way of studying how context affects individuals. This approach offers to be less sensitive to modelling decisions, thereby offering stronger inferences, in situations with strong confounding and/or a lack of common support, as is so often the case when studying neighborhood effects.

The Centrality of Place in Understanding Police Behavior

With important exceptions to be discussed below, both academic and lay explanations of police behavior tend to emphasize the importance of individual and incident-level features (Klahm & Tillyer, 2010; Klinger, 2004). This includes things like what the suspect was doing at the time of an encounter, but by far most of the attention has gone to the role of suspects' race. Particularly, research has a tendency to fixate of whether similarly situated individuals of different races are treated differently by the police, and on the mental processes—like implicit bias—which might explain why this is the case (Kohler-Hausmann, 2018; Neil & Winship, 2019; Russell-Brown, 2018). Whether police engage in this form of differential treatment is a surprisingly difficult question to answer well (Neil & Winship, 2019; Ridgeway & MacDonald, 2010), although in the context of SQF there are some good attempts to do so (Coviello & Persico, 2015; Gelman et al., 2007; Goel et al., 2016; Pierson, Corbett-Davies, & Goel, 2017). However, this emphasis on whether officers are biased against individuals of certain races is misplaced. Even if differential treatment is present, by itself it is unlikely to do a good job of explaining patterns of police behavior. This is certainly the case with SQF, which I will demonstrate in three ways.

First, Goel and colleagues (2016) find that most racial disparities in hit rates disappear after conditioning on precinct. That is, while they conclude that the NYPD engaged in differential treatment in making weapon stops, they also attribute much of the racial disparity in those stops to differences in how police behave in different places. The same conclusion can be drawn from Coviello and Persico (2015), who find that hit rates are not significantly different by race and ethnicity after conditioning on precinct.[1] Ridgeway's (2007) results indicate that much of the racial disparity in SQF stops disappears when conditioning on precinct. Thus, even if all we care about is why racial disparities exist, which is only a part of the larger task of explaining why NYPD officers stopped the people they did, then a focus on individuals alone is misplaced: place matters. However, place is not an explanation in itself (Fagan, 2010), which raises a central question: what is it about place that matters? Below I return to this question at length.

[1] This study differs from Goel et al (2016) in several ways, including the time period studied, the way hit rates are calculated, the specific types of stops examined, and the metric of a hit used.

Second, central to the similarly situated notion is the idea that officers come across people of different races and ethnicities while on patrol and that we should seek to understand whether they treat these people the same (when they're the same in other ways too). While this has an appealing moral aspect to it, in a city as segregated as New York, it is empirically absurd.[2] Figure 1 shows the concentration of SQF weapon stops in New York's neighborhoods, where the cumulative distribution of those stops is compared to the cumulative share of New York's non-Hispanic white residents that neighborhoods contain. For example, 50% of weapon stops occurred in neighborhoods that contained 2.2% of New York's white residents and 80% happened in those with 6.6% of its white residents. If the red line touched the top-left corner, it would indicate that no weapon stops happened in a neighborhood where a white person lived.[3] Reality is surpringly close to that hypothetical: most stops happen where few, if any, white people are found.[4] This suggests that asking whether the police treat similarly situated people of different races the same will not get us very far in explaining either the number of stops or disparities therein; to do so, we need to know why stops are so concentrated in certain areas.

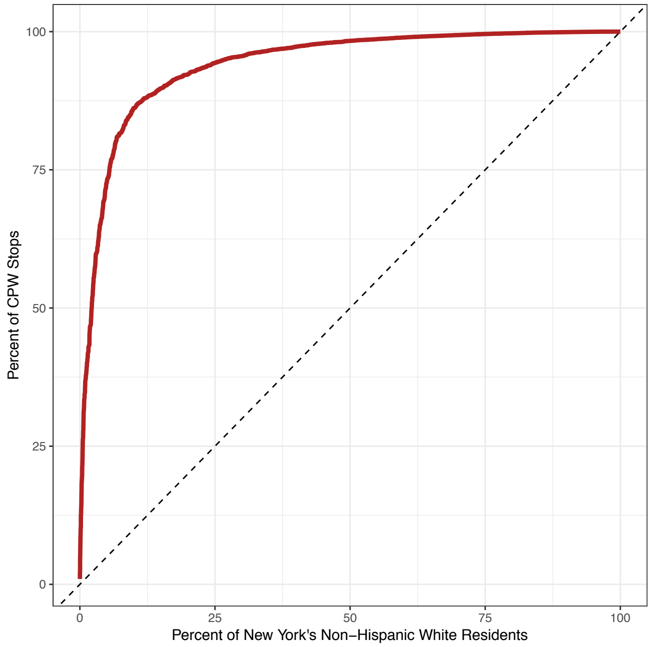

Figure 1: Concentration of Weapon Stops in Tracts, according to fraction of New York's non-Hispanic white population.

[2] See Kohler Hausmann (2018) for a different reason to think that the similarly situated notion is absurd.

[3] The activity space of people is not necessarily the same as their place of residence. But white people don't tend to venture into the black and/or Hispanic neighborhoods where these stops are concentrated. Walking around these neighborhoods, I've seen few non-Hispanic white people.

[4] This is pattern is strongest for weapon stops, the focus of this article, but still holds true for all SQF stops.

Third, while Figure 1 shows massive spatial variation in the number of stops, the typical standard of suspicion that police apply in deciding to make a stop also varies massively by place. This can be seen in Figure 2, which plots the hit rate (specifically, the weapon recovery rate for weapon stops) by both the racial composition of tracts and the race/ethnicity of stopped individuals.[5] In areas where there are more black and/or Hispanic residents, the typical standard of suspicion applied in making stops is much lower. Importantly, this is not simply because there are more black and/or Hispanic individuals stopped in those places (see Fagan et al., 2010 and Goel et al., 2016 for further evidence). Rather, while hit rates are higher for non-Hispanic whites than the comparison groups in all neighborhoods, they consistently fall for all groups the more black and/or Hispanic an area becomes. Hit rates are lower for non-Hispanic white individuals in the most black and/or Hispanic areas than they are for non-Hispanic black or Hispanic individuals in the least black and/or Hispanic areas. Again, these patterns indicate that

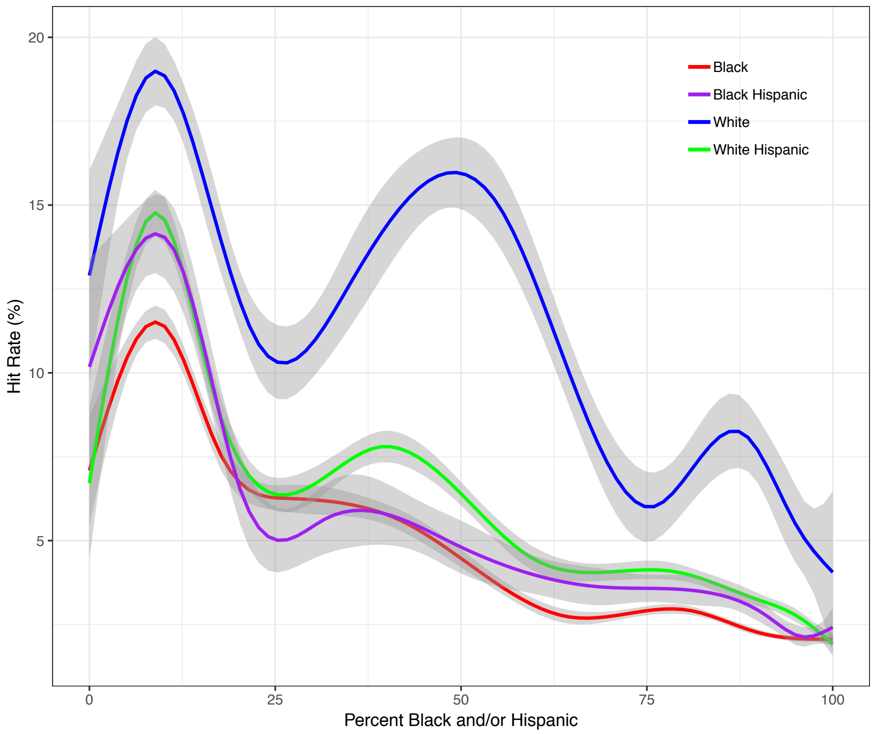

Figure 2: Weapon Recovery Rates of CPW Stops, by Race of Individual and Racial Composition of Tracts.

[5] For reasons that will become apparent, this is an imperfect measure of standards of suspicion. Model 1 of my main result deals with these reasons and arrives at the same general conclusion.

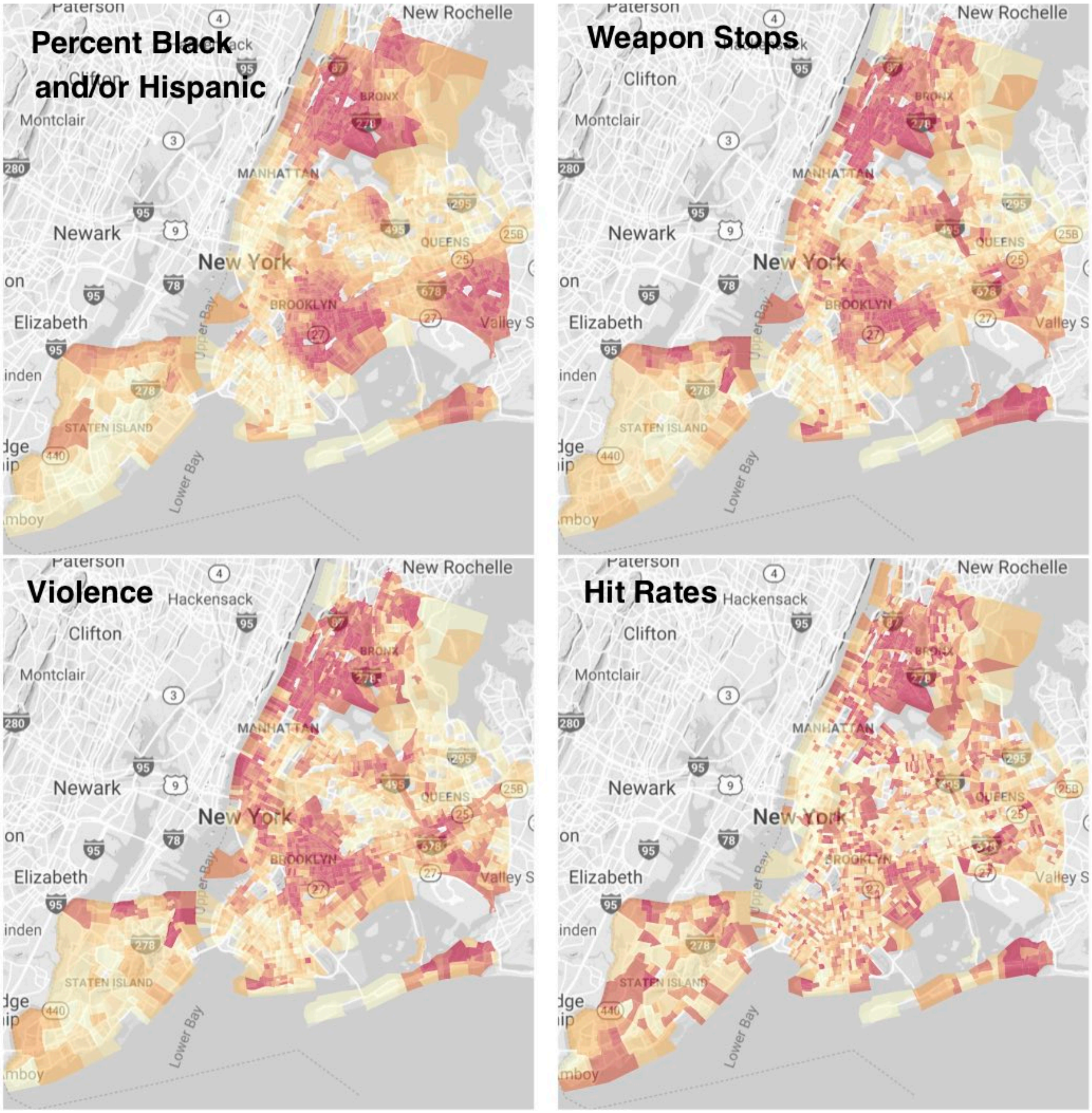

Figure 3: The Distribution of Black/Hispanic Residents, Weapon Stops, Violence, and Hit Rates across Census Tracts, 2008-2012. Each map is shaded according to its decile value, wither darker shades being higher values, except for hit rates where dark indicates lower weapon recovery rates of weapon stops. The hit rates are estimated through a Bayesian multilevel model to smooth the highly-variable estimates. The violence map uses information on the number of violent crimes reported to the NYPD.

understanding why place mattered so much is crucial if we want to understand why the NYPD stopped the people they did. Note that I use percent black and/or Hispanic descriptively here, as a way to sort between different types of neighborhoods. It is not clear why standards of suspicion are so much lower in areas with more of these minority group members, nor why stop counts are so much higher. Figure 3 makes this clear: there is spatial variation in weapon stops and hit rates that aligns very closely with the racial composition of areas, but from visual inspection alone it seems to align just as closely with the amount of violent crime reported to the police.[6] In order to get beyond descriptive differences to understand why this neighborhood variation exists, more sophisticated analyses are needed.

To be clear, these points should not be taken as an argument that individual and incident level factors are unimportant in explaining police behavior. Nor are they evidence against the existence of differential treatment in SQF. Rather, the point is that if we want to understand police behavior—including but not limited to the racial disparities that result from it—by focusing on these things alone we are conditioning on much of the variation that matters. More specifically, for the case under consideration: in order to understand the stops patterns that resulted from SQF, it is necessary to understand why NYPD officers behaved so differently in different neighborhoods.

Ecological Discrimination

While it is an obvious, empirical truth that minority communities are policed more aggressively, my central argument is that they are policed more aggressively because they are minority communities. That is, the racial composition of a neighborhood—in particular the extent to which its residents are black and/or Hispanic—affects how aggressively police behave, including the number of stops made and the typical standards of suspicion applied in deciding to make those stops. This process will produce different levels of enforcement in different neighborhoods but will not necessarily lead to racial and ethnic disparities within neighborhoods. However, it is

[6] Presumably, the hit rate map does not look as similar as the other maps do to each other because they are measuring a rare event and many neighborhoods have very few stops, meaning the estimates are highly variable. Still, the same general pattern is apparent.

discriminatory on account of the fact that it involves race shaping enforcement in a way that is normatively unacceptable, and because it will produce aggregate racial disparities in who comes into contact with police. I call this process ecological discrimination. I posit that this is both an important driver of police behavior, and that it constitutes a particularly important form of discrimination in policing.

There are four mechanisms as to why the racial composition of neighborhoods might shape police behavior. First, the police may perceive people encountered in black and/or Hispanic neighborhoods as particularly suspicious. Sampson and Raudenbush (2004) argue that neighborhoods can carry stigma, showing racial and economic context to be particularly important influences on the level of perceived disorder in Chicago neighborhoods (see also Quillian & Pager, 2001; Sampson, 2009). Hwang (2016) presents the case of a Philadelphia resident who when asked to draw a map of neighborhoods shaded in a large area and called it the "crime area." Some of the shaded area was high-crime, but more than that the shaded areas reflected where black people reside. Context—including the racial and economic composition of neighborhoods—can shape cognition, and this is turn can perpetuate inequality (Sampson, 2012, 2013). The normal mechanism of stratification advanced is that well-off people will avoid such areas, reproducing spatial disadvantages. But rather than avoid these areas, the police are often the people sent in to deal with the problems created by these larger social processes. This represents a different way in which context might shape cognition and ultimately produce inequality.

Sampson and Raudenbush's (2004) work is inspired in part by Werthman and Piliavin (1967, p.76) who describe what they call ecological contamination:

> Residence in a *neighborhood* is the most general indicator used by the police to select a sample of potential law violators. Many local patrolmen tend to consider *all residents* of

> "bad" neighborhoods rather weakly committed to whatever moral order they make it their business to enforce, and this transforms most of the people who use the streets in these neighborhoods into good candidates for suspicion. [emphasis in original]

Seeing a neighborhood as "bad" might make officers police more aggressively, and racial and economic context are important forces in making people see neighborhoods as "bad." Thus, people encountered in minority-filled and poor areas may be seen as particularly suspicious; additionally, they may be seen as particularly deserving of aggressive policing. Klinger (1997) argues that in high-crime areas people might be seen as less deserving of police services, but if the police see their actions not as a protective service but as a punishment to people who they target his argument can be flipped on its head: (perceived) high-crime places might be seen as particularly deserving of aggressive enforcement. While Sampson and Raudenbush (2004) is in part a critique of broken windows theory, these processes have not been studied much as causes of police behavior (but see Fagan & Davies, 2000; Fagan et al., 2010). Grundwald and Fagan (2019) argue that once they control for neighborhood crime, NYPD officers were still more inclined to designate stops they made in areas with more black residents as having happened in "high crime areas." This finding supports this mechanism, though is not definitive proof as it may also reflect an institutionalized "script" that officers apply more heavily in those neighborhoods to justify stops (Fagan & Geller, 2015).

A second mechanism is the ability of communities to hold officers accountable for overly aggressive enforcement. There are two sources of this ability. The first is in the collective efficacy of the community (Sampson, 2012; Sampson, Raudenbush, & Earls, 1997). While informal and formal social control are often presented as competing explanations as to why crime rates vary across communities (Jacobs, 1961; Sampson, 1986a), informal social control can also be exerted over the agents of formal social control. That is, greater social cohesion and a combined willingness to intervene on the common good might make residents more capable of countering the aggressive policing of their communities compared to socially disorganized communities. The second is that people found in high status areas are more likely to be powerful or to be connected to powerful people (Sampson, 2012). Police officers have a strong desire to avoid disciplinary action (Moskos, 2008; Mummolo, 2018), and for instance, randomly stopping someone who turns out to be a well-connected lawyer's son does not bode well for that desire. The result will be more restraint in places where such people are likely to be found. If officers have a sense of these community-based accountability mechanisms, they are unlikely to act aggressively in the first place. As a result, the underlying capacities of powerful communities are likely to remain latent. With these processes, both officer perceptions and the underlying realities matter, but the perceived link between something like poverty and social disorganization is likely far more accurate than that between race and criminality. In New York, poor black and/or Hispanic areas may be more socially disorganized than many whiter areas, meaning these communities are less likely to have the capacity to counter aggressive policing. A similar point is made by Kane (2002) in his study of the social ecology of police misconduct in New York City.

A third reason that minority areas may be policed more intensely pertains to public space. On the one hand, it has been frequently claimed that a core task of the police is to exert control over space, particularly public spaces (Herbert, 1997; Sampson, 1986b; Stinchcombe, 1963; Werthman & Piliavin, 1967). On the other hand, the way that public space is used differs by the socioeconomic status of places and thus by racial context as well. Specifically, areas of low socioeconomic status tend to have a more active street life, in part because being higher class affords access to private space (Stinchcombe, 1963). For example, while residents of Williamsburg might readily go to rooftop bars where drinks cost more than minimum wage on a warm summer's day, many residents of Brownsville who want the same things cannot afford this privilege. What begins as the same desire—to enjoy warm weather with friends over a few drinks—puts poorer people in the public spaces that police see as a core part of their mandate to control, whether that be a housing project, a stoop, or a corner.[7] Independent of the supply of people in such public spaces, if the police see people in places with more minorities as particularly suspicious, disorderly, or criminal—as per the first mechanism—this could lead them to more aggressively assert control over the public spaces in such areas so as to counter perceived threats to public order and safety, or presumed threats to their territorial claim on the public space. Just as a broken window can mean something different depending on the racial and economic context in which it is embedded (Sampson, 2012), so too can a group of people hanging out on a corner.

A fourth mechanism is racial and economic threat (Carmichael & Kent, 2014; D. Jacobs & O'Brien, 1998; Kane, 2003). Both of these threat theories are premised on the idea that there are groups that threaten the dominant social classes/races, and so law enforcement is used to exert social control over these groups. The word threat is used in a very specific way; in contrast to the mechanism where racial composition can make things seem more disorderly/criminal (and thus threatening), here threat means a challenge to the dominant racial and economic order. Thus, if the NYPD was trying to manage the underclass (Wacquant, 2009), or to perpetuate a system of racial hegemony (Alexander, 2012), it would police minority and/or poor people more aggressively. While a plausible mechanism, the fact that NYPD officers were so intensely targeting blacks in only certain areas, and that in those areas they were mostly targeting teenage

[7] My analysis focuses on how standards of suspicion vary across places, rather than the number of stops, and so this part of this mechanism does not apply as it doesn't necessarily require that the police apply different standards of suspicion across different racial or economic contexts.

boys and young men, makes it seem as if the police were going after people that they (often incorrectly) thought were criminals, rather than trying to impose control over those who threatened the racial or economic order. Similarly, this mechanism seems to be more focused on the race and class of individuals rather than of neighborhoods, which also casts doubt on its importance as a neighborhood mechanism. Unless one thinks this mechanism stems from the police organization rather than the individual officer.

Notice that for all of the above mechanisms, even if individual officers were not directly affected by them, they could still produce more aggressive enforcement. This is because they could influence the number of officers that are allocated to an area (Fagan, 2017b; Kane, 2003), or they could lead the police organization to pressure its officers to police more aggressively in minority areas. City leaders and police commanders are always under pressure to make it look as if they're "doing something" in response to the city's problems, and these pressures find their way downwards to patrol officers. All of the four mechanisms described above give reason to think that the racial composition of areas will influence what police commanders decide the something they want done by their officers is, and how much of it they want done. As suggested above, it makes more sense to think racial and economic threat would affect organizational practices and in turn officer behavior, since police organizations are closely tied to those in power, compared to individual officers who themselves are often working class and/or from minority groups. The point is that even if individual officers are "just doing their job" in a neighborhood in a way that is not directly shaped by the racial or class context, the fact that they are in that neighborhood in the first place, the pressure being put on them by their superiors to behave in a certain way (e.g. at roll call), as well as the norms as to "how things are done around here," could all be shaped by the racial ecology in which their part of the police organization is embedded.

Two further observations about these mechanisms are in order. First, they are not necessarily mutually exclusive. For example, perhaps stronger demands from precinct commanders to make more stops in black and/or Hispanic areas combines with officers' perceptions of those areas as filled with criminal people who lack the ability to resist their aggressive enforcement. Second, some of these mechanisms pertain more to the socioeconomic status of an area than to directly how having many people of a given race in a neighborhood impacts officer behavior. I do not view socioeconomic status as something that confounds racial composition, but rather part of what constitutes what it means for an area to be black and/or Hispanic (Kohler-Hausmann, 2018). Ecological discrimination, like any form of discrimination, is fundamentally a normative concept: it involves deciding what are unfair ways for race to produce its effect on the world. I find it unfair that minorities areas may be subject to more intense policing because they are often lower-class areas, and so I include socioeconomic mechanisms as part of ecological discrimination. That said, acknowledging that others may not agree with this normative decision, I also conduct an analysis which attempts to tease the racial composition and socioeconomic status of areas apart.[8]

How "Percent Black" Studies May Go Wrong

To be sure, a large body of research has examined racial composition in relation to police behavior (e.g. Carmichael & Kent, 2014; Fagan, Geller, Davies, & West, 2010; Jacobs &

[8] Because the racial composition and socioeconomic status of neighborhoods in New York (as in other American cities) are very highly correlated, in addition to the conceptual problem in trying to teases them apart, it is also statistically very difficult to do so.

O'Brien, 1998; MacDonald & Braga, 2018). In these studies, the percentage of a neighborhood that is black is usually an independent variable in a regression model where the outcome is something like the number of police stops.[9] The work of Fagan and colleagues on the NYPD's practice of SQF is a particularly well-known subset of this research (Fagan, 2010; Fagan & Davies, 2000; Fagan et al., 2010; Geller & Fagan, 2010). This body of research often discusses race in relation to the ecology of policing, and I have drawn upon it extensively to develop the ideas presented above. However, there are three reasons to doubt the findings of percent black studies.

First, a percent black "effect" is used as evidence of three different things: that racial composition affects police behavior (the focus of this article); that the police engage in disparate treatment along racial lines; or of a total race effect that combines both of these. The problem is that a percent black parameter should not be used as evidence of the latter two things, and if used as evidence of the former it requires modelling competing explanations in a way that has not been done. The implication is that results from percent black studies are likely consistent with a wide range of competing interpretations.

Figure 4 presents a directed acyclic graph (DAG) to elucidate this argument.[10] In this DAG, Y is some police behavioral outcome of interest (e.g. stops) and RC is racial composition, correctly estimating the relationship indicated by the arrow between RC and Y is thus the central task. To do so, there can be no unblocked backdoor paths from RC to Y (Morgan & Winship,

[9] I call studies relating racial composition to police behavior "percent black studies" in this section, since that is nearly always the focal independent variable used, though the point applies more broadly to the relationship between racial composition and police behavior, including to my metric which combines percent black and/or Hispanic.

[10] In this DAG, I do not mean to imply race has a causal effect on police behavior, but rather that the race of individuals might lead officers to treat them differently.

2014). Put differently, it is necessary to control for the race of individuals that the police encounter (RI), and for neighborhood-level (Z) and individual/incident-level (X) features that are related to both RC and Y. As such, a percent black study which attempts to estimate whether racial composition affects police behavior by only measuring neighborhood-level confounders (Z) will likely be biased. This will be the case if individuals and incident-level features that impact police behavior—including the race of individuals—varies in a way that is correlated with racial composition. These lower-level processes (RI and X) need to be controlled for. Notice a practical challenge in doing so: this seems to imply we need information on police encounters that do not lead to stops, such as an officer seeing someone and doing nothing else. Such data does not currently exist and might require an undesirable amount of surveillance of the public to collect. Fortunately, hit rate tests can circumvent this challenge when used properly, as I discuss below.

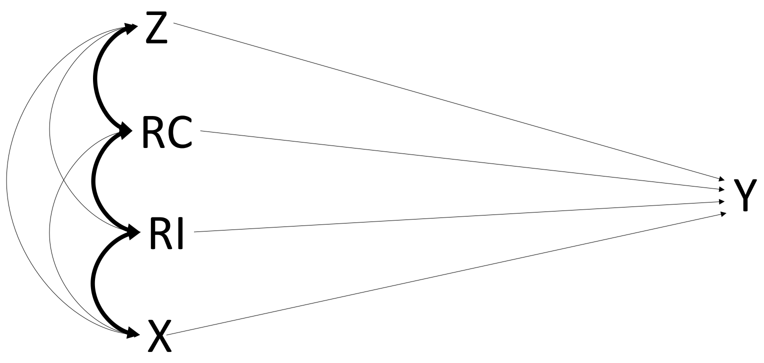

Figure 4: Directed Acyclic Graph (DAG) Displaying Challenges in Estimating Racial Composition Effect on Police Behavior. Y is some police behavioral outcome, RC is racial composition, RI is the race of the individual, X is some set of individual and incident-level features, and Z is some set of neighborhood-level features. Bold arrows are used to indicate what are likely highly correlated relationships.

Models that include RC and Z have been used as a test of differential treatment by the police (Fagan, 2010), what Neil and Winship (2019) call the ecological test. Note that, using the terms from Figure 4, the ecological test estimates the arrow from RI to Y by measuring RC instead. One particularly important problem with this approach is that it omits X, that is, it does not control for differences between individuals and incidents that may be correlated with race (Neil & Winship, 2019). This same observation also means that percent black "effects" should not be interpreted as the total effect of race. Studies frequently interpret a percent black coefficient as indicative of either racial discrimination against individuals and/or the racial composition of neighborhoods structuring police behavior (e.g. Fagan et al., 2010), this is what I mean by total effect. But there is no reason to think that omitting RI and X will turn RC into a correct estimate of the total influence of race: the results of such a model could be entirely driven by the omitted variable bias from not including X, and so race might in reality have no effect, at either level. In short, care must be taken with percent black studies about what is trying to be inferred. Some desired quantities, like the total effect of race or differential treatment, do not seem possible to estimate from a percent black study, whereas studies specifically testing for whether race has an ecological effect must disentangle this from the race of individuals encountered and of the appropriate sets of not only neighborhood, but also individual and incident-level confounders.

The second problem is that even if RC is correctly estimated (i.e. RI, X and Z are properly controlled for), then this isn't necessarily evidence of a particular reason that RC matters. Above, I presented four mechanisms through which race can have an ecological effect on police behavior. Standard practice would be to take a percent black "effect" as evidence of just one of those mechanisms. This is a problem because all of the mechanisms make the same prediction about the percent black coefficient; to sort between which mechanisms are operative it is necessary to go beyond looking for a percent black effect. I do not do so in this paper, since I see all of these mechanisms as part of the larger process of ecological discrimination, but anyone making claims about the importance of specific mechanisms should seek to do so. While there is no problem in interpreting a well-estimated percent black effect as evidence of any of these mechanisms, using this as the only test and then making very strong claims about the existence or importance of this specific mechanism on the basis of this test comes dangerously close to affirming the consequent (i.e. if A then B, B therefore A).

The third problem is that the racial composition parameter (RC) is likely to be strongly confounded, as is represented by the bold lines in Figure 4. This is a problem because percent black studies have (to my knowledge) exclusively relied on regression, and because regression does not perform well in the presence of strong confounding. Thus, even if these models contained all of the right control variables, they face two related yet distinct challenges: imbalance and lack of common support (Gelman & Hill, 2006). When a focal independent variable (e.g. RC) is highly correlated with other covariates (e.g. RI and Z) which themselves are correlated with the outcome (Y), the resulting imbalance will tend to produce regression estimates that are highly sensitive to model specification. A lack of common support would occur, for example, if there were black neighborhoods for which no comparable white neighborhoods existed, where comparability is defined according to values of the other covariates. King and Zeng (2006) explain that in such a situation the extreme counterfactuals being asked require extrapolation beyond the data, thereby making estimates very sensitive to modelling decisions, a situation they term model dependence.

Thus, in the presence of strongly imbalanced data and/or lack of complete overlap, regression—including HLM—becomes very sensitive to model assumptions. Yet neighborhoods, certainly the neighborhoods of American cities like New York, exhibit serious imbalance and lack of overlap along many dimensions, including things that could be important drivers of police behavior like the distribution of serious crimes. Ridgeway (2006) makes a similar argument in the context of post-stop studies focused on individual race, and more recently has provided an example focused on a percent black coefficient using SQF data, showing that the estimated effect is highly sensitive to modelling decisions (Ridgeway, 2017). This is a challenge that goes far beyond the research question under consideration, and below I will discuss the promise of my method to address it in other contexts.[11]

Taking Stock

Understanding neighborhood-level variation in police behavior is central to understanding the NYPD's practice of SQF. The racial composition of neighborhoods might be an important driver of police behavior. Testing this idea faces several difficulties that have been explained in this section. To account for them, my research design must: measure racial composition separately from the race of individuals; control for other aspects of neighborhood that might matter for police behavior and for other incident and individual-level features that might explain away a racial composition-police behavior relationship; and avoid using regression because it is not trustworthy in this context. Below, I implement a matching approach with which to test for ecological discrimination that addresses these challenges head-on.[12]

METHODS

Data

[11] Many of these issues also apply to post-stop studies, although they often do a better job of dealing with the issues discussed in this section. Moreover, ecology may matter most for who gets stopped in the first place, and stops might be the most important part in determining who experiences certain post-stop outcomes, so it is most important to understand the ecology of stops in particular.

[12] In my analysis, I focus solely on the decision-making process, not on the number of decisions (and closely related, decision-makers), even though that is profoundly ecological, as the probability of detection is an important part of who gets stopped (Black & Reiss, 1970).

Data are drawn from three sources. First, the NYPD's Stop, Question, and Frisk dataset supplies a wide array of individual and incident-level information. This publicly-available data is composed of information from UF-250 forms, which are filled by officers when they conduct Terry stops. I use data from the years 2008 to 2012—the height of Stop, Question, and Frisk—and restrict analysis to the roughly one-quarter of the stops in which the suspected crime was criminal possession of a weapon (CPW). Second, spatial crime data from 2007 to 2012 was obtained from the NYPD's publicly available crime complaint dataset.[13] This dataset contains geocoded data on all felony, misdemeanor, and violation crimes reported to the NYPD. While reported crimes are an imperfect measure of actual crime, this is a strength when studying how the police respond to crime, as they cannot respond to the crime that they do not know happened. Finally, spatial information on neighborhoods' physical and social structure was obtained from the American Community Survey (ACS) 5-year estimates (2008-2012).

All but one of the ecological variables are measured at the census tract-level. Tracts are quite small in New York City, which can help incorporate the fact that police structure their behavior along rather small spatial resolutions (Weisburd, Telep, & Lawton, 2014).[14] However, they are large enough so as to be less susceptible to the measurement error which might come from the variance of ACS estimates or from minor errors in the NYPD's geocoding of crimes. Moreover, using smaller spatial resolutions (e.g. block groups) leads to many stops being coded as occurring in unpopulated areas, which upon closer inspection happened in very small

[13] Data from 2007 was only used to provide information on the recent crime levels around stops that occurred in 2008.

[14] Nevertheless, it is unlikely that tract-level measures can fully account for whether police were engaging in hot-spot policing in response to severe violence. This is why I also measure whether stops happened near a recent homicide.

unpopulated areas of what are actually highly-populated neighborhoods. For these stops, tracts more likely reflect the context along which officers structure their behavior.

The focal independent variable uses information on the percent of a tract's residents that are black and/or Hispanic. For reasons discussed below, I discretize this into quintiles.[15] I opt for percent black and/or Hispanic over the more conventional use of percent black (or of examining percent black and percent Hispanic separately), because it more accurately reflects the racialized context in which police officers are operating. Imagine a neighborhood that was 50% black and 50% non-Hispanic white, and one which was 50% black and 50% Hispanic white. Using percent black as the metric, these neighborhoods would be measured equivalently. I posit that for every mechanism described above, the latter hypothetical neighborhood would be more similar to a 100% black neighborhood than to the former. Further information on the variables used in analyses, including which data source each variable is from and how it is operationalized, can be found in Table 1.

[15] According to where each tract's value falls in the tract-level distribution, as opposed to where it falls on the stop incident-level distribution, which is skewed heavily toward tracts with many blacks and Hispanic residents.

Table 1: Variables Used in Analyses

Variable	Description	Used in Which Analyses	Balance Imposed
Data Source: NYPD UF-250 Form			
Weapon Recovered	Binary: yes or no	Outcome Variable	N/A
Suspect Race	Factor: white; black; Hispanic; Native American; Asian; other	1	F
Suspect Sex	Binary: male or female	1	F
Suspect Age	Factor: quintiles	1	F
Suspect Build	Factor: heavy; medium; muscular; thin; unknown	1	F
Suspect Height	Integer: inches	1	M
Suspect Weight	Integer: pounds	1	M
Time of Day	Factor: 12am-4am; 4-8am; 8am-12pm; 12pm-4pm; 4pm-8pm; 8pm-12am	1	F
Year	Integer: years	1	F
Stop Location	Factor: public housing; transit; neither	3	F
Stopped Inside	Binary: inside or outside	1	F
Stopped because: Furtive Movements; Casing; Fit Description; Suspicious Bulge; Acting as Lookout; Clothes; Objects; Drug Transaction; Violent Crime; Other	Ten separate binary variables, each indicating whether or not the officer recorded one of the listed items as a reason for making the stop	1	F
Additional circumstance: Evasive Response; Time Period; Victim/Witness Report; Ongoing Investigation; Associating with Criminal; Sights and Sounds of Crime; Change Direction from Officer; Other	Eight separate binary variables, each indicating whether or not the officer recorded one of the listed items as an additional circumstance they considered when making the stop	1	F
Radio Run	Binary: yes or no	1	F
Observation Period before Stop	Integer: minutes	1	M
Data Source: NYPD Crime Data			
Recent Nearby Murder	Binary: was there a murder within a 500-meter radius in the past 30 days	2	F
Violence	Factor: quintiles, using data on number of incidents at tract-level from previous year	2	F
Property	Factor: quintiles, using data on number of incidents at tract-level from previous year	2	F
Data Source: American Community Survey (all measured at Census tract-level)			
Black and/or Hispanic Composition	Factor: quintiles based off percentage of total population that is black and/or Hispanic	Focal Independent Variable	N/A
Concentrated Disadvantage	Factor: quintiles based off of first component from a principal component analysis using median household income and percentage unemployed, using SSI, families below poverty line, female headed households, less than high school education, bachelor's degree or more	3	F
Young Population	Factor: quintiles based off percentage of total population less than 18	3	F
Household Size	Continuous: average household size	3	M
Density	Continuous: population per sq. mile of land	3	M

Note: In the third column, the numbers indicate which models the variable was used in: 1 = the model that removes individual and incident level features; 2 = the model that adds crime controls; 3 = the model that adds class controls. All variables used in the earlier analyses (i.e. lower numbers) are used in subsequent analyses (those with higher numbers). In fourth column, M = balanced on first moment (mean) and F = fine balance.

The dependent variable is whether a stop recovered a weapon. As such, the outcome being analyzed is a hit rate, the rate at which stops recover weapons. Differences in hit rates across some type of strata (e.g. across racial/ethnic groups or, in my case, spatial contexts) can be taken as evidence that the police apply different standards of suspicion. A lower hit rate for a stratum suggests that people who are unlikely to have a weapon—people who are less suspicious—are being stopped at higher rates in that stratum relative to the comparison stratum. However, two classes of problems threaten to make this interpretation of hit rates incorrect. The first is omitted variable bias and infra-marginality (Ayres, 2002; Dharmapala & Ross, 2004; Neil & Winship, 2019; Ridgeway & MacDonald, 2010). If the police apply a lower level of suspicion due to some other factors, and these other factors are correlated with membership in the strata of interest, then the association may be spurious. Relatedly, if weapon carrying rates differ, the hit rate might be higher for a stratum even if the police apply the same or a lower standard of suspicion to that stratum relative to a comparison stratum. This is a specific instance of what is usually called the infra-marginality problem (Ayres, 2002). To guard against these possibilities, it is necessary to compare the hit rates of stops that are similar except in how they differ in one-dimension, in my case how they differ by racial composition (Neil & Winship, 2019).[16] Uncoincidentally, this is the same conclusion that was arrived at when I presented the DAG above, and it is what my analysis is designed to do.

[16] This may not fully solve the infra-marginality problem. Even in a world where the relationship between racial composition and weapon recovery was unconfounded, including that the circumstances of encounters and characteristics of encountered individuals were extremely similar, weapon carrying rates still might differ by place. If true, and if we are willing to assume weapon carrying rates would be higher in areas with more black and/or Hispanic residents (perhaps due to proximity to high-crime areas), then my results are understating the effect of racial composition. This is because this would bias the hit rate upwards in areas with many black and/or Hispanic residents, where my conclusion is based off of the fact that hit rates are so low in those areas.

The second class of problems pertains to what constitutes a hit. It must be something that is well-measured and that the police were trying to recover with their stops (Goel et al., 2016). If this isn't the case, it isn't clear what hit rates mean, if anything. It is for this reason that I restrict my analysis to weapon stops. Whether a weapon was found is an objective criterion with a direct link to the stated reason of a CPW stop (Goel et al., 2016; Mummolo, 2018). This means that for CPW stops the rate of weapon recovery is both well-measured and particularly apt for gauging whether stops achieve their goal. Fortunately, CPW is by far the most common suspected offense type, meaning my inferences apply to a large fraction of the stops, although I will discuss how this sample restriction limits my conclusions below.

I use listwise deletion as missing data is quite rare. Specifically, of the 760,502 stops eligible for inclusion, there is data on every variable for 714,627 (94%). Another concern is the CPW stops that went unreported. One reason that this is likely not a large problem is that it is much easier and less risky to avoid making a stop than to make it and fail to document it (Mummolo 2018). Not making a stop is less risky because it does not involve approaching someone an officer thinks is armed, and because failing to document a stop could lead to disciplinary action. In addition, particularly in the time period studied, NYPD officers were under intense pressure to make and document stops (Eterno, Barrow, & Silverman, 2017; Mummolo, 2018; Rayman, 2010). Recovering a weapon, especially a gun, was highly-regarded within the department (Mummolo, 2018), and so presumably an officer documenting that they were making CPW stops was also particularly well-received by supervisors. Still, in about one-fifth of the investigated SQF-related complaints in 2012, officers did not fill in the mandatory UF-250 form (Schneiderman, 2013). To the extent to which the forms that are not filled are a random subset of the CPW stops, this issue will not bias findings. During the time period I study, the NYPD had audits in place to ensure the accuracy of filled UF-250 forms (Ridgeway, 2007). While the data is imperfect, I assume it is complete and accurate enough so as to not distort findings.

Matching with a Multi-Valued Treatment and Hierarchical Data

Matching is a means of ameliorating the challenges posed by strongly imbalanced data and/or a lack of common support (Ho, Imai, King, & Stuart, 2007). There are many different forms of matching, and they do not tend to perform equally well (King, Nielsen, Coberley, Pope, & Wells, 2011). I use cardinality matching, which directly targets balance via optimization (Visconti & Zubizarreta, 2018; Zubizarreta, Paredes, & Rosenbaum, 2014). That is, it turns matching into an optimization problem, where the goal is to maximize pair-matched sample size subject to user-defined balance constraints. This is a very flexible approach as there are many possible choices of balance constraints, such as balancing on various moments of marginal or joint distributions, exact matching, and fine balance (Visconti & Zubizarreta, 2018). More formally, cardinality matching solves the problem:

$$\max_{m} \sum_{t \in T} \sum_{c \in C} m_{tc}, \tag{1}$$

$$\sum_{c \in C} m_{tc} \leq 1, \;\; t \in T, \tag{2}$$

$$\sum_{t \in T} m_{tc} \leq 1, \;\; c \in C, \tag{3}$$

$$\left| \sum_{t \in T} \sum_{c \in C} m_{tc} \left(f_k\left(x_{tp}\right) - f_k\left(x_{cp}\right) \right) \right| \leq \varepsilon_p \sum_{t \in T} \sum_{c \in C} m_{tc}, \tag{4}$$

where T represents the set of observations in the treatment condition and t indexes specific observations in that set and (equivalently) C represents the set of control observations and c specific observations, p indexes the observed covariates x, and m_{tc} is a binary decision variable which equals 1 if a treated unit is matched to a control unit, and 0 otherwise (Visconti & Zubizarreta, 2018). Thus, Equation 1 maximizes sample size, subject to constraints in Equations

2 and 3 which enforce the formation of a pair-matched sample, and to the balance constraints of Equation 4. In Equation 4, f_k is a function that transforms the observed covariates for each balance condition $k \in K$, and $\varepsilon_p \geq 0$ is a tolerance. Both the balance conditions and tolerance are defined by the researcher. The end result will be the largest pair-matched sample that meets these user-imposed balance constraints.

Three features of my study require that this cardinality matching approach is modified. One is that I want my quantity of interest to be representative of a well-defined population, specifically I want to estimate an average treatment effect (ATE). Neither regression (Aronow & Samii, 2016) nor the pair-matching procedure described above will necessarily produce an ATE, even if the estimated treatment effects are unbiased and consistent. Another is that matching across multiple levels of treatment (or equivalently "exposure") is NP-hard: it is not computationally tractable (Bennett, Vielma, & Zubizarreta, 2018). These two challenges are dealt with by a recent extension to cardinality matching proposed by Bennett and colleagues (2018). They propose first creating a template, that is, a sample of observations that is representative of a target-population. Then, the observations of each level of treatment are matched to this template. Since all levels of treatment are balanced compared to the same template, they will also be balanced to compared to each other. Moreover, because this template represents a meaningful population, an estimate obtained by comparing levels of treatment will also represent a meaningful, well-defined quantity of interest.

In order to estimate an ATE, my template is a random sample of 4,000 observations from the entire dataset (of 714,627 observation). Since randomization only produces balance in expectation, I drew 500 such samples of 4,000, compared how similar each was to the entire dataset in terms of Mahalanobis distance, and selected that which was most similar to the entire dataset as my template.[17] I discretized the percent of a tract's residents who are black and/or Hispanic into quintiles, because this approach cannot handle a truly continuous treatment variable.[18] Each of these five levels is matched to the template. I impose two forms of balance constraints: mean balance and fine balance. Fine balance perfectly balances the marginal distribution of a categorical (or ordinal) variable (Rosenbaum, Ross, & Silber, 2007). This can be used directly on categorical variables; additionally, discretizing continuous variables and imposing fine balance is an excellent way to guarantee certain segments of the distribution are balanced across treatment levels. The type of balance constraint imposed for each variable can be found in Table 1.

The final required modification stems from that fact that because many stops happen in the same places, their errors are not independent. Measures of uncertainty must account for this dependent error structure or they will be biased towards zero. To do so, I employ cluster-robust bootstrapping (Field & Welsh, 2007). While conventional bootstrapping samples unit-level observations with replacement as a non-parametric means of estimating sampling distributions, this procedure does not preserve dependence in the error structure. To circumvent this, cluster-robust bootstrapping samples at the cluster-level instead.

My approach is certainly more complicated than HLM. It is worth remembering the payoff for this effort: in contexts with heavily imbalanced data and/or a lack of common support—as is often the case when studying neighborhood effects—regression results are likely driven by assumption. By non-parametrically balancing the data, the matching approach addresses these challenges head-on, so that we may compare unit-level outcomes for

[17] Closer examination of this template compared to the entire dataset revealed that they are indeed extremely similar.

[18] The cut points are 13.32%, 29.92%, 68.70%, and 91.67% black and/or Hispanic.

observations that differ only by one aspect of the cluster in which they are embedded. This is exactly what is needed to estimate the racial composition parameter shown in the DAG of Figure 4. More broadly, this approach offers stronger inferences about how context affects individuals, including but not limited to neighborhood effects.

Plan of Analysis

Analysis proceeds in four steps. First, I compare stops that are the same in terms of individual and incident-level features, but which occurred in neighborhoods that differed in terms of their racial and ethnic composition. Second, I repeat this analysis but also equalize neighborhood crime conditions. Then, I repeat that analysis but equalize key features of neighborhood socioeconomic status. Finally, supplementary analyses ensure that results are not overly sensitive to decisions about the data and models used.

RESULTS

Comparing Similar Stops in Different Areas

The first step—Model 1—compares the hit rates of stops that are similar in terms of pertinent individual and incident features (see Table 1 for what these are) but differ by the ecological contexts in which they occur. In Figure 4, this model blocks RI and X. This should be thought of as answering the descriptive question: how much do standards of suspicion vary by ecological contexts, net of incident and individual-level features? The answer, as is visualized in the top row of Figure 5, is that they vary massively. Figure 5(a) shows the hit rate across neighborhoods, where the hit rate varies from 7.08% to 2.73% in those neighborhoods with the fewest black and/or Hispanic residents to those with the most, respectively. Figure 5(b) uses this same data but presents it instead as a contrast compared to Level 1 (the least black and/or Hispanic neighborhoods). These same results are presented in Table 2. As can be seen there, the differences between Level 1 compared to Levels 3, 4, and 5 are all significant (p<0.05). Notably, differences of this magnitude in hit rates are as large or larger than those that have been reported in studies looking for individual-level discrimination (Coviello & Persico, 2015; Goel et al., 2016). Put differently, standards of suspicion vary massively across neighborhoods that differ by racial composition. This is better evidence of the argument that was made earlier on: in order to understand SQF it is necessary to understand why this neighborhood-level variation exists. Before attributing these observed differences to an effect of racial composition, it is necessary to rule of alternative explanations (the Zs in Figure 4).

Table 2: Results from Matching, Contrasts with Level 1

	Model 1: Controlling for Individual and Incident Variables	**Model 2: Adding Neighborhood Crime Controls**	**Model 3: Adding Neighborhood Class Controls**
Level 2	-1.13 (-3.06, 0.86)	-0.23 (-3.91, 3.13)	-0.49 (-3.25, 1.89)
Level 3	-2.75* (-4.62, -0.98)	-2.39 (-6.12, 0.58)	-1.32 (-3.73, 1.27)
Level 4	-3.45* (-5.13, -1.84)	-3.37* (-7.10, -0.17)	-2.08* (-4.48, 0.00)
Level 5	-4.35* (-6.04, -2.81)	-4.87* (-8.35, -2.12)	-2.99* (-5.39, -0.66)

*Notes: Level 5 is that with highest percentage black and/or Hispanic residents. Brackets contain 95% bootstrap confidence intervals for the contrast with Level 1. * = p<0.05.*

Controlling for Crime

Model 2 controls for the level of property and violent crime, and for whether there has been a recent, nearby homicide (see Table 1 for further details). That the NYPD applies a lower standard of suspicion when making weapon stops in areas that have higher-levels of crime, particularly violent crime, is not something I consider to be ecological discrimination, but rather a legitimate alternative explanation for observed neighborhood-level differences in hit rates (even if, ultimately, SQF is not a particularly effective weapon-recovery strategy). As is seen in

Figure 5(c), sharp differences remain by area. The hit rates now vary from 7.82% to 2.95%, indicating that the police applying a lower standard of suspicion in making weapon stops as areas become more black and/or Hispanic. Figure 5(d) presents this same data as contrasts compared to Level 1; these contrasts are also presented in Table 2. The contrasts between Level 1 compared to Levels 4 and 5 are significant (p<0.05). It is not clear what could produce these observed differences except the four mechanisms of ecological discrimination described above.

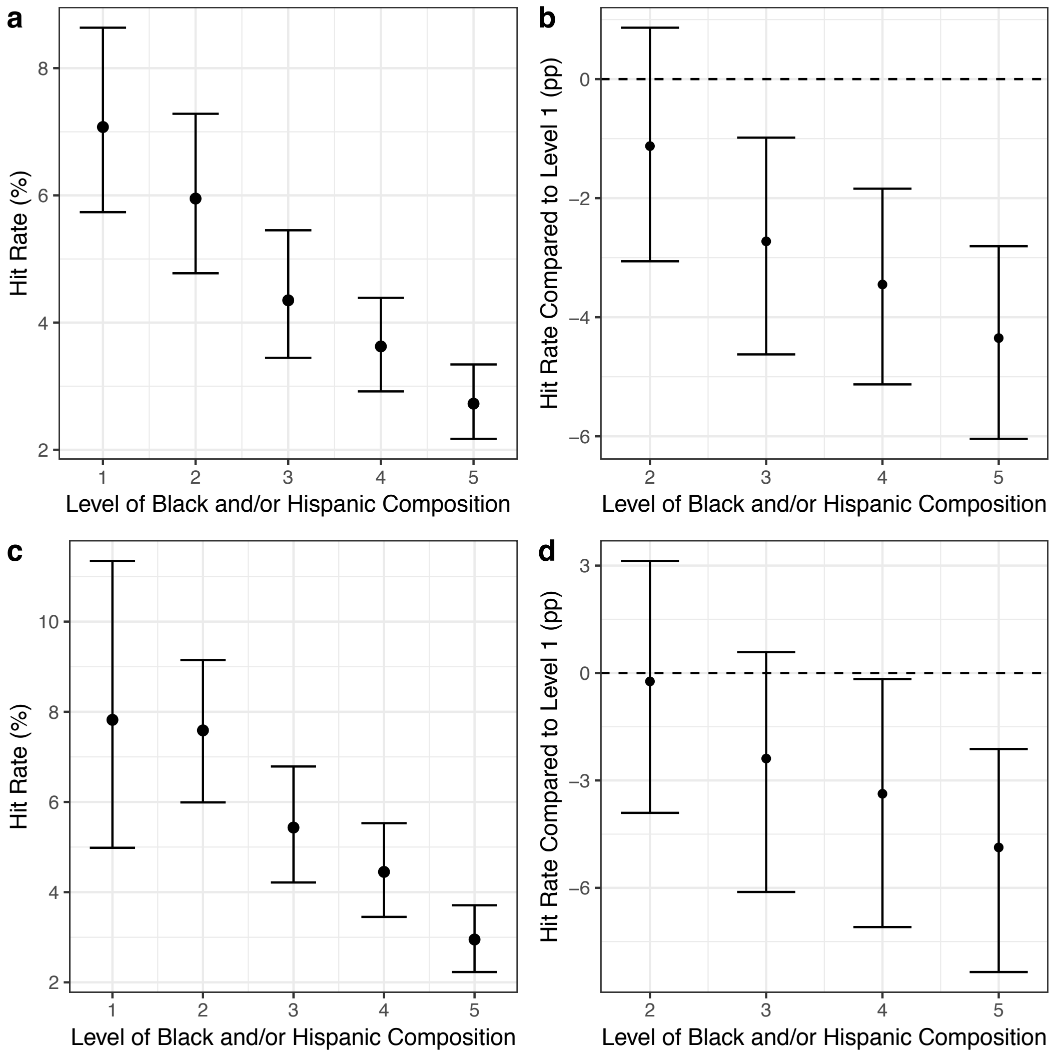

Figure 5: Results from Models 1 and 2. (a) The mean hit rate by racial composition for Model 1. (b) The contrast with Level 1 for the other levels of Model 1. (c) The mean hit rate by racial composition for Model 2. (d) The contrast with Level 1 for the other levels of Model 2. Bars indicate 95% confidence intervals. Level 5 is those tracts with highest composition of black and/or Hispanic residents.

Controlling for Neighborhood Socioeconomic Status

As I argue above, neighborhood class is best conceptualized as a mechanism of ecological discrimination rather than a confounder. Nevertheless, in this section I attempt to parse the two apart, despite the conceptual problems and empirical challenges in doing so. I do so by controlling for concentrated disadvantage, and also for household size, population density, the age composition of the population, and whether stops happened in public housing, transit, or neither (see Table 1 for more details on these measures). After doing this, the hit rates vary from 6.67% to 3.68%. The contrasts with Level 1 can be found in Table 2: the differences between that and Levels 4 and 5 are significant ($p<0.05$). Thus, the observed differences in hit rates is smaller than in earlier analyses, but still present. I take this as evidence that black and/or Hispanic areas are not policed more aggressively solely because they tend to be of lower socioeconomic status, but also because of more purely racial mechanisms of ecological discrimination.

Supplementary Analyses

In Appendix A, Table 3 presents results from a sensitivity analysis. These are E-values, which indicate the minimum strength of association between both tract racial/ethnic composition and an unmeasured confounder and between the outcome and an unmeasured confounder that would be required to explain away the difference in hit rates across areas defined by racial composition, on the risk ratio scale (VanderWeele & Ding, 2017). For all comparisons, Level 1 is treated as the reference category. It appears that results are moderately robust to unmeasured confounding, particularly the contrasts between Level 1 and the most black and/or Hispanic areas. In Appendix B, a plot is used to assess covariate balance. For parsimony, I only present the results from Model 2, but one arrives at the same conclusion when examining Models 1 and 3 in the same way: as it is intended to do, the cardinality matching has produced excellent balance on the covariates across levels of the treatment. Finally, using HLM to adjust for remaining imbalances in Models 2 and 3 keeps the contrasts between Level 1 and Level 5 significant ($p<0.05$).

DISCUSSION

In summary, in areas with more black and/or Hispanic residents, the NYPD officers tend to require a lower standard of suspicion when deciding to make weapons stops. This is not simply a descriptive difference across areas. Even after accounting for the race of stopped individuals, other pertinent features of those individuals and the stop incidents, and neighborhood crime rates, hit rates still varied massively across areas. The racial composition of neighborhoods itself affects police officer behavior. I have not identified the specific reason, as four processes seem possible: that the racial context impacts cognition, that social disorganization allows for more aggressive policing, that public spaces are used and policed differently in areas of lower socioeconomic status, or that racial threat is operative. That racial composition still has an effect after equalizing neighborhood socioeconomic status strongly imply that racial threat or how racial composition impacts officer cognition are part of the explanation. Because the main effect is a hit rate, it is somewhat hard to interpret, but it is a large effect. Indeed, the differences in hit rates due to racial composition may be larger than the differences in hit rates due to the race of individuals (Goel et al., 2016). In short, to make sense of NYPD stop patterns and racial disparities therein, it is necessary to understand why officers behave so differently in different neighborhoods. I have found the racial composition of neighborhoods to be an important part of this process.

While explaining Stop, Question, and Frisk is an important task in its own right, my findings offer three sets of broader implication. First, by focusing on individual and incident-level features, including whether officers discriminate on the basis of individual's race and the associated mental biases which might explain why they do so, extant work may be effectively conditioning on most of the variation that it seeks to explain and thus missing important driving forces of police behavior. As I have demonstrated in the context of the NYPD's SQF, place matters massively in explaining what police do. But place is not an explanation in itself, we must ask what it is about place that matters. In this vein, I have advanced the concept of ecological discrimination, the idea that the police discriminate at the level of the neighborhood. Specifically, I have argued neighborhoods that are more black and/or Hispanic may end up being policed more aggressively, and there are four possible mechanisms explaining why this is the case. Ecological discrimination appears to be an important factor in explaining spatial variation in the standards of suspicion that NYPD officers applied when making stops. This suggests that those seeking to understand police behavior in other contexts would be well-served by trying to understand the spatial variation in police behavior, and that ecological discrimination should be taken as a likely important candidate explanation.

A second implication of my research is that future research which studies racial composition in relation to police behavior needs to be very careful about what exactly it is claiming and whether the models used are a good test of that claim. I have shown that earlier such work (what I call "percent-black studies") is not sufficiently attentive to whether race is supposed to have an individual-level effect, an ecological effect, or some combination of both, that it does not do enough to control for confounders at both the individual/incident-level and ecological levels, and that consequently it is not clear what the findings of such studies mean.

Additionally, even if one is careful to control for these competing explanations so as to isolate an ecological race effect, it could still have many explanations. In my case, any of the four mechanisms are possible. Should researchers desire to study the existence or importance of one of these specific mechanisms, they will need to avoid simply relying on interpreting a percent black coefficient—as is standard practice—because all mechanisms make the same prediction so doing that cannot discern between them.

A final broad implication of my research is that scholars need to seriously question the use of regression, including HLM, when making causal inferences about how context affects individuals, including inferences about neighborhood effects. If some treatment variable is highly confounded with other variables, or if other covariates do not share common support across treatment level, then regression will likely produce results that are highly sensitive to modelling assumptions. Yet in neighborhoods, certainly in America but also elsewhere, things go together: many features of neighborhoods tend to be very closely related, to the point where the different neighborhoods might as well be different worlds, or to be less hyperbolic, to the point that there are few to no good counterfactuals across neighborhoods defined by something like their racial composition. In these circumstances, data can only get us so far. Yet it is possible to do better than regression. To that end, I have adapted a novel use of cardinality matching in order to study how context affects individuals. In principal, this method should produce inferences about neighborhood affects that are less likely driven by assumption. Given that understanding how context affects individuals is a key part of the sociological enterprise, many scholars should find this approach useful for their specific research problems.

The main limitation of my analysis, an issue with all analyses using NYPD SQF data, is that I do not know which individuals are stopped multiple times (Neil & Winship, 2019). If the police were intensely targeting a small subset of individuals, this would likely bias the racial composition parameter estimate. In the language of Figure 4, this is a potentially important X that is omitted from the model. Still, sensitivity analyses have indicated that the results are moderately robust to the omission of such a variable. That I have only examined weapon stops means that I have not answered whether ecological discrimination was operative for other types of SQF stops. As such, the extent to which my findings apply more generally requires assumptions. My assumption is that ecological discrimination was operative for other types of SQF stops; I say this because nothing about the four mechanisms discussed gives any reason to think that if they existed, they would only affect weapon stop patterns.

I conclude with a general observation. While contemporary sociological research has begun to present evidence on the experience and consequences of being policed, it has been largely silent on the causes of police behavior (Brayne, 2014; Goffman, 2014; Legewie & Fagan, 2019; Rios, 2011; Stuart, 2016). This is a surprising state of affairs given that policing is a fundamental social institution in the modern world, because of the highly charged debates surrounding police behavior in American society, and because of the police's role as the front end of the criminal justice system which has drawn so much attention from sociologists over the past two decades. As this paper illustrates, due to its emphasis on taking evidence seriously and not reducing human behavior to the product of decontextualized cognitive states or preferences, a sociological perspective can offer important insights on why the police do what they do. Long ago, this was the case (Bittner, 1970; Black & Reiss, 1970; Smith, 1986; Werthman & Piliavin, 1967). Now that policing and criminal justice are such central concerns, it should be so once again.

Bibliography

Alexander, M. (2012). *The New Jim Crow: Mass Incarceration in the Age of Colorblindness*. The New Press.

Aronow, P. M., & Samii, C. (2016). Does Regression Produce Representative Estimates of Causal Effects? *American Journal of Political Science*, *60*(1), 250–267. https://doi.org/10.1111/ajps.12185

Ayres, I. (2002). Outcome Tests of Racial Disparities in Police Practices. *Justice Research and Policy*, *4*(1–2), 131–142. https://doi.org/10.3818/JRP.4.1.2002.131

Bennett, M., Vielma, J. P., & Zubizarreta, J. R. (2018). Building Representative Matched Samples with Multi-valued Treatments in Large Observational Studies: Analysis of the Impact of an Earthquake on Educational Attainment. *ArXiv:1810.06707 [Stat]*. Retrieved from http://arxiv.org/abs/1810.06707

Bittner, E. (1970). *The functions of the police in modern society; a review of background factors, current practices, and possible role models.* Chevy Chase: National Institute of Mental Health, Center for Studies of Crime and Delinquency.

Black, D. J., & Reiss, A. J. (1970). Police Control of Juveniles. *American Sociological Review*, *35*(1), 63–77. https://doi.org/10.2307/2093853

Borchetta, J. R., Charney, D., & Harris, A. (2018, April 11). Don’t Let the Police Wreck Stop-and-Frisk Reforms. *The New York Times*. Retrieved from https://www.nytimes.com/2018/04/10/opinion/police-stop-and-frisk-reforms.html

Brayne, S. (2014). Surveillance and System Avoidance Criminal Justice Contact and Institutional Attachment. *American Sociological Review*, *79*(3), 367–391. https://doi.org/10.1177/0003122414530398

Carmichael, J. T., & Kent, S. L. (2014). The Persistent Significance of Racial and Economic Inequality on the Size of Municipal Police Forces in the United States, 1980–2010. *Social Problems*, *61*(2), 259–282. https://doi.org/10.1525/sp.2014.12213

Coviello, D., & Persico, N. (2015). An Economic Analysis of Black-White Disparities in the New York Police Department’s Stop-and-Frisk Program. *The Journal of Legal Studies*, *44*(2), 315–360. https://doi.org/10.1086/684292

Dharmapala, D., & Ross, S. L. (2004). Racial Bias in Motor Vehicle Searches: Additional Theory and Evidence. *Contributions in Economic Analysis & Policy*, *3*(1). https://doi.org/10.2202/1538-0645.1310

Eterno, J. A., Barrow, C. S., & Silverman, E. B. (2017). Forcible Stops: Police and Citizens Speak Out. *Public Administration Review*, *77*(2), 181–192. https://doi.org/10.1111/puar.12684

Fagan, J. (2010). *Report of the court, Floyd et al. v. city of New York, 08 Civ 01034 (SAS). U.S. District Court for the Southern District of New York*.

Fagan, J. (2017a). Recent Evidence and Controversies in "the New Policing." *Journal of Policy Analysis and Management*, *36*(3), 690–700. https://doi.org/10.1002/pam.21995

Fagan, J. (2017b). Response to Ridgeway: Allocating Police. *Journal of Policy Analysis and Management*, *36*(3), 703–707. https://doi.org/10.1002/pam.21998

Fagan, J., & Davies, G. (2000). Street Stops and Broken Windows: Terry, Race, and Disorder in New York City. *Fordham Urban Law Journal*, *28*, 457–504.

Fagan, J., & Geller, A. (2015). Following the Script: Narratives of Suspicion in Terry Stops in Street Policing Symposium: Criminal Procedure in the Spotlight. *University of Chicago Law Review*, *82*, 51–88.

Fagan, J., Geller, A., Davies, G., & West, V. (2010). Street stops and broken windows revisited: The demography and logic of proactive policing in a safe and changing city. In *Race, Ethnicity, and Policing: New and Essential Readings* (pp. 309–348). New York: New York University Press.

Field, C. A., & Welsh, A. H. (2007). Bootstrapping clustered data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *69*(3), 369–390. https://doi.org/10.1111/j.1467-9868.2007.00593.x

Floyd et al. v. City of New York, No. 08 Civ. 1034 (SAS) (Dist. Court August 12, 2013).

Geller, A., & Fagan, J. (2010). Pot as Pretext: Marijuana, Race, and the New Disorder in New York City Street Policing. *Journal of Empirical Legal Studies*, *7*(4), 591–633. https://doi.org/10.1111/j.1740-1461.2010.01190.x

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. https://doi.org/10.1198/016214506000001040

Gelman, A., & Hill, J. (2006). *Data Analysis Using Regression And Multilevel/Hierarchical Models*. Cambridge: Cambridge Univ. Press.

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. https://doi.org/10.1214/15-AOAS897

Goffman, A. (2014). *On the Run: Fugitive Life in an American City*. Macmillan.

Grundwald, B., & Fagan, J. (2019). The End of Intuition-Based High-Crime Areas. *California Law Review*, *forthcoming*.

Herbert, S. K. (1997). *Policing Space: Territoriality and the Los Angeles Police Department*. U of Minnesota Press.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis*, *15*(3), 199–236.

Hwang, J. (2016). The Social Construction of a Gentrifying Neighborhood: Reifying and Redefining Identity and Boundaries in Inequality. *Urban Affairs Review*, *52*(1), 98–128. https://doi.org/10.1177/1078087415570643

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. https://doi.org/10.1086/231291

Jacobs, J. (1961). *The Death and Life of Great American Cities*. Macat Library.

Kane, R. J. (2002). The Social Ecology of Police Misconduct. *Criminology*, *40*(4), 867.

Kane, R. J. (2003). Social control in the metropolis: A community-level examination of the minority group-threat hypothesis. *Justice Quarterly*, *20*(2), 265–295. https://doi.org/10.1080/07418820300095531

King, G., Nielsen, R., Coberley, C., Pope, J. E., & Wells, A. (2011). *Comparative Effectiveness of Matching Methods for Causal Inference*. Working Paper, IQSS, Harvard University.

King, G., & Zeng, L. (2006). The Dangers of Extreme Counterfactuals. *Political Analysis*, *14*(2), 131–159. https://doi.org/10.1093/pan/mpj004

Klahm, C. F., & Tillyer, R. (2010). Understanding police use of force: A review of the evidence. *Southwest Journal of Criminal Justice*, *7*(2), 214–239.

Klinger, D. A. (1997). Negotiating Order in Patrol Work: An Ecological Theory of Police Response to Deviance. *Criminology*, *35*(2), 277–306. https://doi.org/10.1111/j.1745-9125.1997.tb00877.x

Klinger, D. A. (2004). Environment and Organization: Reviving a Perspective on the Police. *The Annals of the American Academy of Political and Social Science*, *593*, 119–136.

Kohler-Hausmann, I. (2018). *Eddie Murphy and the Dangers of Counterfactual Causal Thinking About Detecting Racial Discrimination* (Working Paper). New Haven, CT: Yale Univesity.

Legewie, J., & Fagan, J. (2019). Aggressive Policing and the Educational Performance of Minority Youth. *American Sociological Review*, 0003122419826020. https://doi.org/10.1177/0003122419826020

MacDonald, J., & Braga, A. A. (2018). Did Post-Floyd et al. Reforms Reduce Racial Disparities in NYPD Stop, Question, and Frisk Practices? An Exploratory Analysis Using External and Internal Benchmarks. *Justice Quarterly*, *In press*.

Morgan, S. L., & Winship, C. (2014). *Counterfactuals and Causal Inference* (2nd ed.). Cambridge: Cambridge University Press.

Moskos, P. (2008). *Cop in the Hood: My Year Policing Baltimore's Eastern District*. Princeton University Press.

Mummolo, J. (2018). Modern Police Tactics, Police-Citizen Interactions, and the Prospects for Reform. *The Journal of Politics*, *80*(1), 1–15. https://doi.org/10.1086/694393

Neil, R., & Winship, C. (2019). Methodological Challenges and Opportunities in Testing for Racial Discrimination in Policing. *Annual Review of Criminology*, *2*.

Pierson, E., Corbett-Davies, S., & Goel, S. (2017). Fast Threshold Tests for Detecting Discrimination. *ArXiv:1702.08536 [Cs, Stat]*. Retrieved from http://arxiv.org/abs/1702.08536

Quillian, L., & Pager, D. (2001). Black Neighbors, Higher Crime? The Role of Racial Stereotypes in Evaluations of Neighborhood Crime. *American Journal of Sociology*, *107*(3), 717–767. https://doi.org/10.1086/338938

Rayman, G. (2010). The NYPD Tapes: Inside Bed-Stuy's 81st Precinct. Retrieved March 21, 2018, from https://www.villagevoice.com/2010/05/04/the-nypd-tapes-inside-bed-stuys-81st-precinct/

Ridgeway, G. (2006). Assessing the Effect of Race Bias in Post-traffic Stop Outcomes Using Propensity Scores. *Journal of Quantitative Criminology*, *22*(1), 1–29. https://doi.org/10.1007/s10940-005-9000-9

Ridgeway, G. (2007). *Analysis of Racial Disparities in the New York Police Department's Stop, Question, and Frisk Practices*. Santa Monica: RAND Corporation.

Ridgeway, G. (2017). Stop-and-Frisk Is Essential … and Requires Restraint. *Journal of Policy Analysis and Management*, *36*(3), 683–689. https://doi.org/10.1002/pam.21990

Ridgeway, G., & MacDonald, J. M. (2010). Methods for Assessing Racially Biased Policing. In S. K. Rice & M. D. White (Eds.), *Race, Ethnicity, and Policing* (pp. 180–204). New York: NYU Press. https://doi.org/10.18574/nyu/9780814776155.003.0007

Rios, V. M. (2011). *Punished: Policing the Lives of Black and Latino Boys*. NYU Press.

Rosenbaum, P. R., Ross, R. N., & Silber, J. H. (2007). Minimum Distance Matched Sampling With Fine Balance in an Observational Study of Treatment for Ovarian Cancer. *Journal of the American Statistical Association*, *102*(477), 75–83. https://doi.org/10.1198/016214506000001059

Russell-Brown, K. (2018). The Academic Swoon over Implicit Racial Bias: Costs, Benefits, and Other Considerations. *Du Bois Review: Social Science Research on Race*, *15*(1), 185–193. https://doi.org/10.1017/S1742058X18000073

Sampson, R. J. (1986a). Crime in Cities: The Effects of Formal and Informal Social Control. *Crime and Justice*, *8*, 271–311. https://doi.org/10.1086/449125

Sampson, R. J. (1986b). Effects of Socioeconomic Context on Official Reaction to Juvenile Delinquency. *American Sociological Review*, *51*(6), 876–885. https://doi.org/10.2307/2095373

Sampson, R. J. (2009). Disparity and diversity in the contemporary city: social (dis)order revisited1. *The British Journal of Sociology*, *60*(1), 1–31. https://doi.org/10.1111/j.1468-4446.2009.01211.x

Sampson, R. J. (2012). *Great American City: Chicago and the Enduring Neighborhood Effect*. Chicago: University of Chicago Press.

Sampson, R. J. (2013). The Place of Context: A Theory and Strategy for Criminology's Hard Problems. *Criminology*, *51*(1), 1–31. https://doi.org/10.1111/1745-9125.12002

Sampson, R. J., & Raudenbush, S. W. (2004). Seeing Disorder: Neighborhood Stigma and the Social Construction of "Broken Windows." *Social Psychology Quarterly*, *67*(4), 319–342. https://doi.org/10.1177/019027250406700401

Sampson, R. J., Raudenbush, S. W., & Earls, F. (1997). Neighborhoods and Violent Crime: A Multilevel Study of Collective Efficacy. *Science*, *277*(5328), 918–924. https://doi.org/10.1126/science.277.5328.918

Schneiderman, E. T. (2013). *A Report on Arrests Arising from the New York Police Department's Stop-and-Frisk Practices*. Albany: Office of the New York State Attorney General.

Smith, D. A. (1986). The Neighborhood Context of Police Behavior. *Crime and Justice*, *8*, 313–341.

Stinchcombe, A. L. (1963). Institutions of Privacy in the Determination of Police Administrative Practice. *American Journal of Sociology*, *69*(2), 150–160.

Stuart, F. (2016). *Down, Out, and Under Arrest: Policing and Everyday Life in Skid Row*. University of Chicago Press.

VanderWeele, T. J., & Ding, P. (2017). Sensitivity Analysis in Observational Research: Introducing the E-Value. *Annals of Internal Medicine*, *167*(4), 268–274. https://doi.org/10.7326/M16-2607

Visconti, G., & Zubizarreta, J. R. (2018). Handling Limited Overlap in Observational Studies with Cardinality Matching. *Observational Studies*, 33.

Wacquant, L. (2009). *Punishing the Poor: The Neoliberal Government of Social Insecurity*. Duke University Press.

Weisburd, D., Telep, C. W., & Lawton, B. A. (2014). Could Innovations in Policing have Contributed to the New York City Crime Drop even in a Period of Declining Police Strength?: The Case of Stop, Question and Frisk as a Hot Spots Policing Strategy. *Justice Quarterly*, *31*(1), 129–153. https://doi.org/10.1080/07418825.2012.754920

Werthman, C., & Piliavin, I. (1967). Gang Members and the Police. In D. J. Bordua (Ed.), *The police: Six sociological essays*. New York: John Wiley and Sons.

Zimring, F. E. (2011). *The City That Became Safe: New York's Lessons for Urban Crime and Its Control*. Oxford: Oxford University Press.

Zubizarreta, J. R., Paredes, R. D., & Rosenbaum, P. R. (2014). Matching for balance, pairing for heterogeneity in an observational study of the effectiveness of for-profit and not-for-profit high schools in Chile. *The Annals of Applied Statistics*, *8*(1), 204–231. https://doi.org/10.1214/13-AOAS713

Appendix A

Table 3: E-Values Needed to Explain Away Difference in Hit Rates Compared to Level 1

	Model 1: Individual and Incident Variables	**Model 2: Adding Neighborhood Crime Variables**	**Model 3: Adding Neighborhood Class Variables**
Level 2	1.66	1.21	1.37
Level 3	2.64	2.24	1.80
Level 4	3.31	2.91	2.27
Level 5	4.63	4.74	3.02

Note: E-values are for the point estimates. Level 1 is the reference category, that is, areas with the lowest percentage of black and/or Hispanic populations.

Appendix B

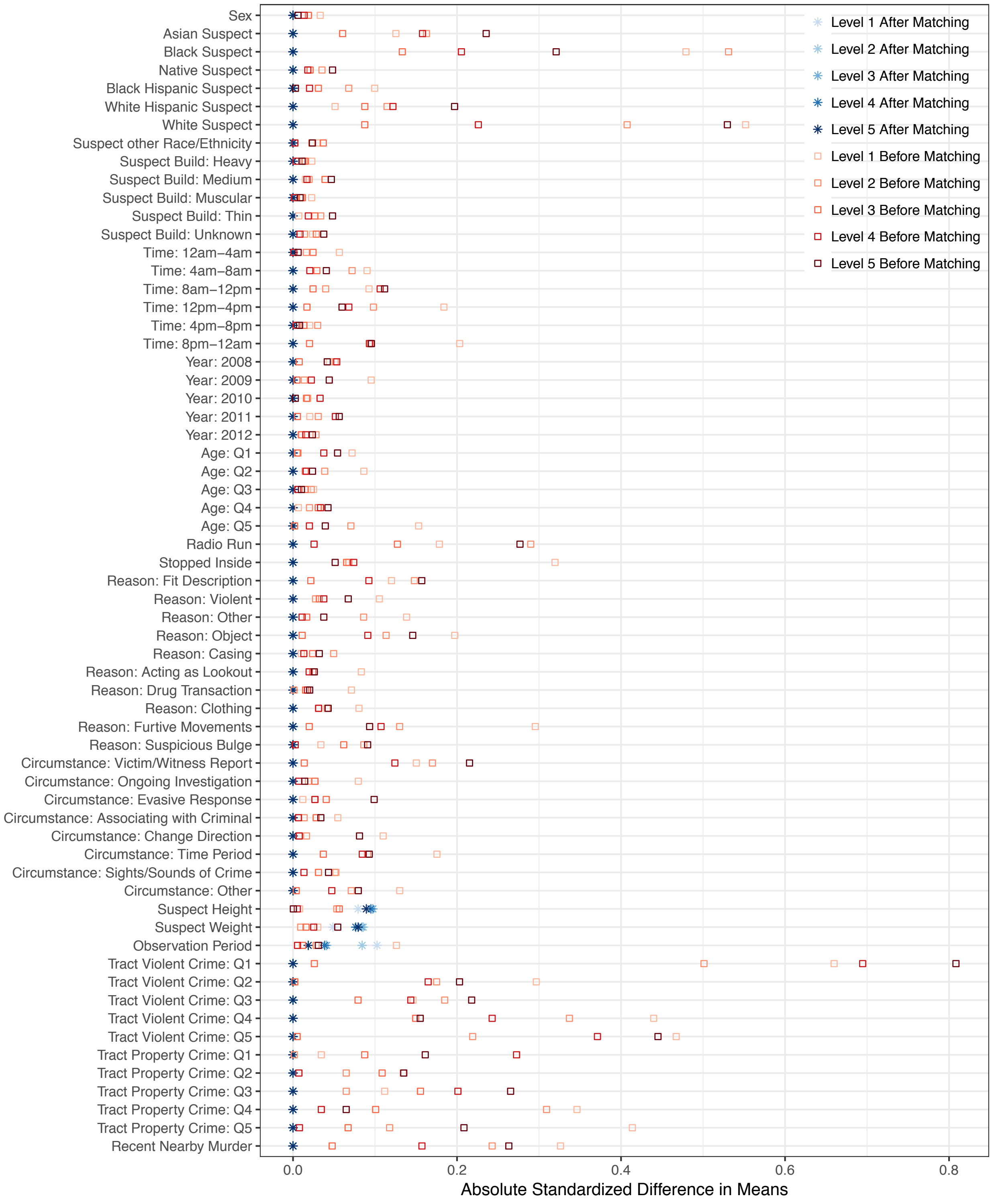

Figure 6: Love Plot Exhibiting Balance of Model 2. The tight clustering around 0 of the matched results (blue shades) indicates that cardinality matching has produced excellent balance for all levels of black and/or Hispanic composition. This is not the case for the unmatched data (red).