Appeared as Chapter 7 in S. Rice and M. White (eds.), 2010, Race, Ethnicity, and Policing: New and Essential Readings (NYU Press) pp. 180–204

Methods for Assessing Racially Biased Policing

Greg Ridgeway
RAND Corporation

John MacDonald
University of Pennsylvania

Abstract

As part of the response to allegations of racially biased police practices many police agencies began collecting information on the stops made by their officers. Social scientists have attempted to use these administrative data on stop decisions to assess the existence or extent of racially biased policing and, in the process, have developed a number of benchmarks for comparison to police stop data. This chapter describes an array of benchmarking methods that have been used around the country including the use of U.S. Census population estimates, non-at fault driver crash data, crime and arrest data, drivers' license data, red light cameras, observations, instrumental variables, assessments of post-stop outcomes, and officer-to-officer comparison via internal benchmarks. Each method's application, strengths, and weaknesses are discussed in the context of their ability to establish a reasonable estimate of the population at risk for being stopped by the police and to draw a causal inference about the extent to which race is a relevant factor in police decision-making on whom to stop, question, and search.

INTRODUCTION

Over the past ten years there has been a proliferation of research that has attempted to estimate the level of racial bias in police behavior. Many police agencies now mandate that their officers record official contacts made with citizens during routine traffic or pedestrian stops. These administrative data sources typically include a host of information on characteristics of the stops made by police officers including: the race/ethnicity of the driver or pedestrian; reasons for the stop; and the actions that occurred after the stop, such as searches, contraband found, and citations or arrests made. These data have been the source for the majority of studies of racially biased police behavior. Analysts have sought to apply basic social science methods to assess whether police agencies as a whole, or in some cases individual police officers, are acting in a racially biased manner. A consistent theme in this research is the search for the appropriate benchmark[i] for which one can quantitatively assess whether police behavior is conducted in a racially biased manner. Studies have linked police administrative data on stops made by officers to a variety of data sources including; police arrest data, population estimates collected by the Bureau of the Census; drivers license data; motor vehicle traffic accident data, moving violations data, systematic observations of drivers, and other sources. Analysts have also attempted to estimate racial bias from assessments of post-stop outcomes and examinations of the "hit rate" (contraband found) from searches. Post-stop outcomes have also focused on matching strategies to appropriately compare minorities and whites that were similarly situated. More recently, efforts have been made to assess individual police officer bias by peer-group officer comparisons.

In the following sections we outline the various methods that have been employed in studies of racially biased policing. We provide an overview of the use of external benchmarks, internal benchmarks, and post-stop outcomes analysis for assessing racial profiling. Our discussion is not an exhaustive review of the literature. Rather, we focus on assessing the methods, their appeal, and there substantive limitations. Developing an appropriate benchmark is more complicated than is presumed in media reports. All of the methods we review for assessing racially biased policing have weaknesses, but some approaches are clearly stronger than others. There is no unifying method that can be applied to administrative data sources and definitively answer the question of whether the police are acting with racial bias. A key issue we address is the fact that the majority approaches used do not meet the basic bedrock assumptions necessary for drawing a causal inference about the effect of race on police behavior. Yet, over time the methods have improved and the policy discussions have inevitably become more nuanced and productive leading to discussions about what the police should and should not be using as pretexts for their decisions on whom to stop and question.

EXTERNAL BENCHMARKS

There is a compulsion in media reports on racial disparities in police stops to compare the racial distribution of the stops to the racial distribution for the community's population as estimated by the US Census. For example, in 2006 in New York City, 53% of stops police made of pedestrians involved black pedestrians while according to the US Census they comprise only 24% of the city's residential population. When the two racial distributions do not align, and they

[i] This is sometimes referred to as the denominator from the standpoint that the proportion of minority stops should be divided by the population at risk (e.g., % black stops/% blacks at risk for being stopped) to provide an appropriate adjustment for detecting racial disparities.

seem to do so rarely, such statistics promote the conclusion that there is evidence of racial bias in police decision making. Racial bias could be a factor in generating such disparities, but a basic introductory research methods course in the social sciences would argue that other explanations may be contributing factors. For example, differences by race in the exposure to the police and/or the rates of committing offenses may also contribute to racial disparities in police stop decisions. It is well documented, for example, that due to historical differences in racial segregation, housing tenure, poverty, and other sociopolitical factors minorities in the US are more likely to live in neighborhoods with higher rates of crime and disorder[ii]. Police deployment in many cities also corresponds to differences in the demand for police services. Neighborhoods with higher volumes of calls to the police service typically have a higher presence of police[iii]. Additionally, research indicates that racial minorities, and in particular blacks, are disproportionately involved in serious personal offenses as both victims and offenders.[iv]

The crux of the external benchmarking analysis is to develop a benchmark that estimates the racial distribution of the individuals who would be stopped if the police were racially unbiased and then comparing that benchmark to the observed racial distribution of stopped citizens. The external benchmark can be thought of as the population at risk for official police contact. As we will see, estimating the appropriate population at risk is complicated. Crude approximations of the population at risk for police contact are poor substitutes and can hide evidence o racial bias or lead to exaggerated estimates of racial bias.

The racial composition of the stops made by the police involves some combination of police exposure to offending/suspicious activity, the racial distribution of the population involved in those activities, and the potential for racial bias. To provide some context, we use some hypothetical numbers and consider an unbiased officer on a foot post who makes stops only when a pedestrian matches a known suspect description. This officer works in a precinct with 40 blacks matching suspect descriptions and 40 whites matching suspect descriptions. If we could somehow measure such numbers we would be inclined to propose a suspect-description benchmark of 50% black and 50% white. However, if the routine daily activities of whites and blacks differ than the officer will encounter different proportions of suspects by race. Say, for example, that the majority of the 40 white suspects stay inside most of the day, travel only by car, or avoid the specific areas with high

[ii] Sampson, R. and W.J. Wilson. Toward a theory of race, crime, and urban Inequality, pp. 37–54.

[iii] Skogan. Disorder and decline: Crime and the spiral of decay in American neighborhoods.

[iv] Hindelang. Variations in sex-age-race incidence rates of offending, pp. 461-475.

police presence, then this officer will stop only a small number of white suspects, deviating substantially from the 50 percent benchmark. Even the less extreme situation, in which half of the white suspects are exposed to the officer, results in the officer stopping blacks in 67 percent of all of their stops decisions. The suspect benchmark in this context is only valid if the police are equally exposed to suspects from the various racial groups. Therefore, even with unbiased officers, we cannot necessarily expect what seems like a reasonable external benchmark to match the racial distribution of stops. This example effectively demonstrates that any of the external benchmarks described in this section must be viewed with caution.

The primary reason for using US Census data to form the benchmark is that it is inexpensive, quick, and readily available. A number of studies attempting to assess racial bias in police behavior use population data from the census, some rely on estimates at local area levels like neighborhood census tracts (see Parker and Stults in this volume). However, for the reasons previously listed, benchmarking with census data does not help us isolate the effect of racial bias from differential exposure and differential offending. Even refinements to the residential census, such as focusing on subpopulations likeliest to be involved in crime (e.g., men or driving age young adults) are not likely to eliminate differences in the exposure of officers to criminal suspects or provide a good approximation of the population at risk for official police action. Fridell[v] summarized the problem with using the census as a benchmark with regard to offender exposure by noting that, "this method does not address the alternative hypothesis that racial/ethnic groups are not equivalent in the nature and extent of their . . . *law-violating behavior*" (p. 106, emphasis in original).

Census estimates provide only the racial distribution of residents and not how these numbers vary by time of day, business attractors such as shopping centers, daily traffic patterns involving commuters, etc. It is quite conceivable that the residential population in many neighborhoods has little resemblance to the patterns of people on the street during the day or night. Even if refinements in the census to the neighborhood or age-prone population at risk for police involvement could give a racially unbiased estimate of the population at risk for police contact, the differences between the residential population and the population at different times of the day and street segments are likely to overwhelm such an estimate. Commuting patterns, for example, can easily exaggerate the racial disparities in traffic stops. Imagine that 20% of traffic stops in a neighborhood that is 95% non-white are made of white citizens. In this context we would suggest whites are stopped 4 times the rate of their composition of the neighborhood population

[v] Fridell. By the numbers: A guide for analyzing race data from vehicle stops.

(20/5=4) and are subjects of racially biased police behavior. However, the stop rate may be a simple reflection of the fact that daily commuters reflect 20% of drivers in this neighborhood.

Dissatisfaction with the census as a benchmark has led some researchers to develop alternate external sets of benchmarks. Some studies of traffic stops attempt to acquire more precise estimates of the racial distribution of drivers on the road to serve as the external benchmark. Under such an approach, one should be able to compare the race distribution of traffic stops made by the police to the race distributions of drivers on the same roadways. Zingraff and colleagues[vi], for example, used the race distribution of licensed drivers rather than the residential population to estimate the race distribution of drivers at risk of being stopped by the police. Although this approach accounts for racial differences in the rate at which the population holds driver's licenses, it does not account for out-of-jurisdiction drivers or for potential racial differences in travel patterns, driving behavior, or exposure to police. To address the problem with out-of-jurisdiction drivers Farrell and colleagues[vii] borrowed driving population models from the transportation literature, which use an area's ability, based on employment or retail location, to pull drivers in from outside communities or to push residents outside the area. This certainly improves upon the census benchmark. However, it is widely documented that minorities (and even those who possess a driver's license) are more likely to take public transit to work and vary from whites in other important ways in their daily travel patterns. Therefore, a more accurate external benchmark would be one that could reliably take into account equivalent driving patterns and behavior between race groups.

Recognizing these limitations, Alpert and colleagues[viii] used data on the location of traffic accidents and the race of the not at-fault drivers to estimate the race distribution of the at-risk population. The logic of this approach is that the race distribution of not-at-fault drivers should approximate the racial distribution of the population of drivers. Although this approach may measure the race distribution of drivers on the road, it does not account for potential racial differences in driving behavior that may be important sources for police decision-making, such as the likelihood of speeding, weaving through traffic, and driving slower than usual.

[vi] Zingraff, et al. Evaluating North Carolina state highway patrol data.

[vii] Farrell et al. Rhode Island traffic stop statistics act.

[viii] Alpert, et al. Toward a better benchmark: Assessing the utility of not-at-fault traffic crash data in racial profiling research.

Other analysts have studied the race distribution of drivers flagged by photographic stoplight enforcement cameras[ix] and by aerial patrols.[x] The advantage of these benchmarks is that they are truly race-blind and measure some form of traffic violation. One can question whether they capture race differences in other aspects of stop risk, such as seatbelt usage, equipment violations, and the other cues that police use in deciding whether or not to stop a citizen.[xi]

Given that the police are not likely to stop people at random, comparisons of racial distribution of stops to the residential population or the driving population on the roadways tells one very little about the race neutrality of the police. Again, it is necessary to establish a benchmark for the population at risk for official police contact. This means that one needs an accurate estimate of the subpopulation that is likely to elicit reasonable suspicion by the police.

Observation benchmarks

Observation benchmarks are a popular approach for attempting to estimate the subpopulation at risk for police behavior. Observation benchmarks typically involve fielding teams of observers to locations to tally the racial distribution of those observed driving and violating traffic laws. More than three decades ago Albert Reiss Jr. advocated the use of systematic social observation as a key measurement strategy for studying the police and other social phenomena.[xii] By systematic, he meant that the observation of behaviors and recordings are done according to explicit standardized rules that permit replication.

This methodology was pioneered to study racial bias in police traffic stops by Lamberth[xiii] in his study of the New Jersey turnpike. Observation benchmarks greatest potential occurs in its application to racial profiling on freeways, since vehicles have essentially the same exposure to the police and speeding is the primary violation that highway patrol focuses on. Speeding, for example, accounted for 89% of the stop reasons in a subsequent study of New Jersey turnpike traffic stops.[xiv] Measuring speeding through direct observations with radar

[ix] Montgomery County Department of Police, Traffic stop data collection analysis.

[x] McConnell, et al. Race and speeding citations: Comparing speeding citations issued by air traffic officers with those issued by ground traffic officers.

[xi] Alpert, et al. Police suspicion and discretionary decision making during citizen stops, pp. 407-434.

[xii] Reiss, Systematic social observation of natural social phenomena, pp. 3-33.

[xiii] Lamberth, Revised statistical analysis of the incidence of police stops and arrests of black drivers/travelers on the New Jersey Turnpike.

[xiv] Maxfield, R. and G. Kelling. New Jersey State Police and stop data.

guns, for example, provides a standardized approach that is easy to replicate and less subject to measurement error than accounting for other types of traffic violations that require observers to make judgments about infractions like weaving through traffic or making illegal turns. Lang and colleagues[xv] and Alpert and colleagues provide two case studies using radar guns.[xvi] The main wrinkle in the analysis of benchmarks based on observation of speeding is determining the appropriate speed at which drivers should be considered "at-risk" for being stopped in specific sections of the highway. For example, it is conceivable that in some areas the police are more vigilant with speeding. As long as this variation is not confounded with differences in the areas that minorities and whites travel than it can provide an unbiased assessment of racial disparities in highway traffic stops.

In urban environments, however, officers stop vehicles for a variety of reasons beyond simple moving violations. Exposure to police can vary widely across different geographic segments of the city.[xvii] In the current volume the reader will note that a number of authors attempt to take the intra-city variation in exposure to the police into account (see e.g., Fagan and Davies). Eck and colleagues[xviii] note that in the city of Cincinnati the police allocate a greater share of officers to areas with a higher volume of crime incidents, and these areas happen to be comprised of predominantly black residents. Relying on direct observations of traffic violations in different segments of the city of Cincinnati would not provide an unbiased assessment of the population at risk for police exposure, because race is confounded with the areas that police are concentrated. One would have to develop an observation method that appropriately balanced these differences in police resource allocation.

There are few examples where investigators have attempted to take the complexity of geographic areas of a city into account in using observation methods. Alpert and colleagues[xix] provide one of the few published studies where trained observers recorded traffic violations (e.g., illegal turns, running stop lights, speeding) at sixteen high volume intersections in Miami-Dade County in areas that were classified as predominately white, black, or racially mixed. A comparison of the racial distribution of observed traffic violators to actual police traffic stops in the same areas suggested little evidence of racial bias in stop decisions. Even if observers in this study did produce an accurate benchmark for individuals at risk

[xv] Lange, et al. Speed violation survey of the New Jersey turnpike.

[xvi] Alpert, et al., Investigating racial profiling by the Miami-Dade Police Department, pp. 25-56.

[xvii] Smith, The Neighborhood context of police behavior, pp. 313-341.

[xviii] Eck, et al. Vehicle police stops in Cincinnati.

[xix] Alpert, et al., Investigating racial profiling by the Miami-Dade Police Department, pp. 25-56.

for exposure to the police in these areas—a challenge on its own right—several issues remain. There is no reason to believe that police stops should be representative of those simply observed in these areas committing traffic violations. Officers target behaviors that they believe indicate drug transactions, stop individuals fitting suspect descriptions, and respond to calls for service. Once observers head down the path of trying to determine which vehicles or persons should be at-risk for being stopped, the observations become more subjective and less systematic.[xx] In fact, the variation between-observers in such studies can exceed the estimate of the racial disparity. One observer may be more likely than others to measure some driving behavior as aggressive. Such variation in judgments in an observation study has to be taken into account, or observers have to be trained to near uniformity in judgments if one is going to produce a reliable estimate of the population at risk for police contact. Regardless, it is unclear that observational studies are relying on the same sets of markers that the police use in deciding who is suspicious and whom to stop. The courts have not consistently supported the use of observational benchmarks for this reason. In United States v. Alcaraz-Arellano[xxi] the court rejected the benchmark, since it was developed for a general population, not those violating the law.

Outside of traffic stop studies on speeding or moving violations on roadways, systematic observations of driving behavior are not likely to yield useful estimates for an external benchmark for an entire city. Recognizing these limitations a number of investigators have turned to other approaches for establishing external benchmarks.

Arrest and crime suspect benchmarks

Gelman, Fagan, and Kiss[xxii] quote then–NYPD Police Commissioner Howard Safir:

The racial/ethnic distribution of the subjects of stop and frisk reports reflects the demographics of known violent crime suspects as reported by crime victims. Similarly, the demographics of arrestees in violent crimes also correspond with the demographics of known violent crime suspects. (2007, p. 4)

Safir is clearly suggesting that violent crime suspects or violent crime arrestees provide a reasonable benchmark from which the public can judge the department's racial distribution in stop percentages. This quote suggests that the

[xx] Ibid.

[xxi] 302 F. Supp. 2d 1217, 1229–1232, D. Kan., 2004

[xxii] Gelman, et al. An analysis of the New York City Police Department's "stop-and-frisk" policy in the context of claims of racial bias, p*p*. 813–823.

arrestee population may serve as a useable benchmark for assessing racial bias in the police decision for whom to stop.

The arrestee benchmark, however, is also problematic because it is too narrow. For example, the police make stops for trespassing, vandalism, suspected drug sales, and a variety of other causes. Many stop decisions might be made for minor infractions, not serious crime incidents involving violence. The group of individuals stopped by the police in most large cities, therefore, far exceeds the group comprising the arrestee population. There are a variety of reasons that the racial distribution of individuals stopped by the police could have a race distribution that differs greatly from that of arrestees. For one, arrests can often take place some distance away from where the crime actually occurred. Most problematic is that, if officers are in fact racially biased then we cannot use their arrests to represent what we would expect of an unbiased police force. Such a benchmark could actually hide bias. Investigators like Gelman and colleagues have attempted to control for this by using prior year arrest decisions as an external benchmark. Again, there is no reason to expect that previous year decisions are independent of current year decisions – especially if as research by Klinger[xxiii] suggests an established pattern of practices becomes ingrained in specific police precincts.

The criminal suspect benchmark may be more plausible approach than the arrestee benchmark for establishing the population at risk for official police contact. It represents the public's reporting of those involved in suspicious activity and crime and would correspond more closely to racial distribution of criminals on the street.[xxiv] Note that this benchmark is not a reasonable choice for traffic stops since often police have the intent to cite for a traffic violation without the expectation that it will lead to an arrest. Comparing the police to the public's reporting of suspicious activity at least answers the question whether the police are finding suspicious individuals with features similar to those the public reports committing or attempting to commit crimes. Ridgeway, for example, found that in New York City black pedestrians were stopped at a rate 20 to 30 percent lower than their representation among the public's report of crime-suspect descriptions and Hispanic pedestrians were stopped slightly more than their share of crime suspect descriptions, by 5 to 10 percent.[xxv] However, the public may have their

[xxiii] Klinger, Negotiating order in patrol work: An ecological theory of police response to deviance, pp. 277–306.

[xxiv] For a discussion of the benefits and limitations of citizens' calls for police service data see Klinger, D. and G. Bridges. Measurement error in calls-for-service as an indicator of crime, pp. 705-726.

[xxv] Ridgeway. Analysis of racial disparities in the New York Police Department's stop, question, and frisk practices.

own racial biases and they may also under or over-report certain activities (e.g., drug market activity, suspicious individuals) depending on the area and the perceived problems that the police actively target.

Instrumental variables

An ideal scientific method to estimate the extent of race bias in policing would be to use an experimental design and randomly assign police officers to be "race blind" during certain periods. For example, for each officer and for each hour that officer patrols the street we flip a coin to determine whether that officer will be unable to perceive the race of a suspect. The difference between the percentage of stops involving minorities when the officers can perceive race to the percentage of stops involving minorities when the officers are race blind gives us the effect of racial bias. If the officers were unbiased then the ability to perceive race should not matter in the selection of stopped individuals. If instead the officers are racially biased then we would observe more minority stops when the officers are not blinded to race.

Clearly such an experiment in the actual field is a fantasy, but instrumental variables (IV) analysis is an econometric approach that can sometimes solve such problems.[xxvi] Instrumental variables analysis relies on the randomization that occurs in nature to replicate the classic randomized experimental design. They key hurdle is to identify an "instrument," in this case a variable that is predictive of the ability to perceive race[xxvii] that is not related to the actual race of suspects[xxviii]. This is a generalization of the setup in the previous paragraph where our coin is the instrument, highly predictive of the ability to see race but unassociated with the race of potentially stopped individuals.

Grogger and Ridgeway[xxix] proposed as an instrument the natural variation in daylight and darkness that switches with the change in daylight savings. It is associated with the ability to perceive race but is not related to the race of drivers on the road. The randomization in nature that diminishes the ability of officers to view the actual race of suspects during specific times of the year may serve as an effective instrument for assessing racial bias in police traffic stops. Presumably the probability of race being visible is greater in daylight. Besides the logic of the

[xxvi] For technical details see Angrist, et al. Identification of causal effects using instrumental variables, pp. 444-455.

[xxvii] This is known as the nonzero average causal effect of the instrument on actual treatment assignment.

[xxviii] This is known as the exclusion restriction.

[xxix] Grogger, J. and G. Ridgeway. Testing for racial profiling in traffic stops from behind a veil of darkness, pp. 878-887.

statement, there is some evidence from the literature supporting this. Lamberth described a traffic survey in which the driver's race could be identified in 95% of the vehicles, but for which nighttime observations required auxiliary lighting.[xxx] Greenwald canceled plans for evening surveys after his observer could identify the race of only 6% of the drivers viewed around dusk.[xxxi]

The logic of this approach goes back to the work of Neyman[xxxii] in the 1920s and is a special case of more general instrumental variable methods. We first have to difference the percentage of black drivers among those stopped between daylight and the percentage of black drivers stopped during darkness. Second, to account for the fact that sometimes race is not visible during the day and can be visible at night, the difference in the percentage of blacks stopped needs to be divided by the difference in the probability of race being visible in daytime and darkness. Importantly, this estimate does not require complete race blindness at night and complete visibility during the day, only a substantive diminished capacity.

One of the difficulties that Grogger and Ridgeway faced when attempting to estimate this instrumental variable is that there is no direct measure of diminished capacity due to changes in daylight, the second step of the described IV estimator. A controlled scientific experiment could be conducted to estimate visibility by daylight and darkness, but this might not reflect the types of lighting situations that officers commonly experience on the streets, especially in parts of the city that are better lit than others. As a result Grogger and Ridgeway's analysis simply assumed, logically, that the denominator is positive, such that the probability of race being visible is greater in daylight.

The validity of this instrument also depends on race being independent of daylight/darkness visibility. However, the race distribution of drivers on the road and exposed to the police may be quite different between daylight hours and nighttime hours. If there were mostly black drivers on the road at night then the analysis would indicate that officers stop an excessive fraction of black drivers during the night, but this would just be because there are a larger proportion of black drivers on the road at night. To correct this potential confound Grogger and Ridgeway controlled for clock-time and compared stops occurring near the changes to and from daylight savings time. On one Monday stops at 6pm occur in daylight and the following Monday stops at 6pm occur in darkness. If we can assume that the race distribution of drivers on the road at 6pm does not change

[xxx] Lamberth. Racial profiling data analysis study: Final report for the San Antonio Police Department.

[xxxi] Greenwald. Final report: Police vehicle stops in Sacramento, California.

[xxxii] Neyman. On the application of probability theory to agricultural experiments.

with daylight savings time and that the police do not suddenly reallocate their officers, then this provides a valid instrument.

Figure 1 demonstrates the idea using data from the City of Oakland. The horizontal axis indicates the clock time and the vertical axis indicates hours since dark. Throughout the analysis, we omit stops carried out during the roughly 30-minute period between sunset and the end of civil twilight, since that period is difficult to classify as either daylight or dark. The solid points indicate stops of black drivers, whereas open circles represent stops of non-black drivers. At any time between 5:19 and 9:06 pm, some stops are carried out when it is dark (gray shading) and some are carried out when it is light (no shading). The diagonal bands are a result of the natural variation in daylight hours over the course of the study period. In particular, the large diagonal gap is a result of the shift from Pacific Daylight Time to Pacific Standard Time at the end of October. This shift is especially useful for our comparison since it creates extremes in visibility for fixed clock times.

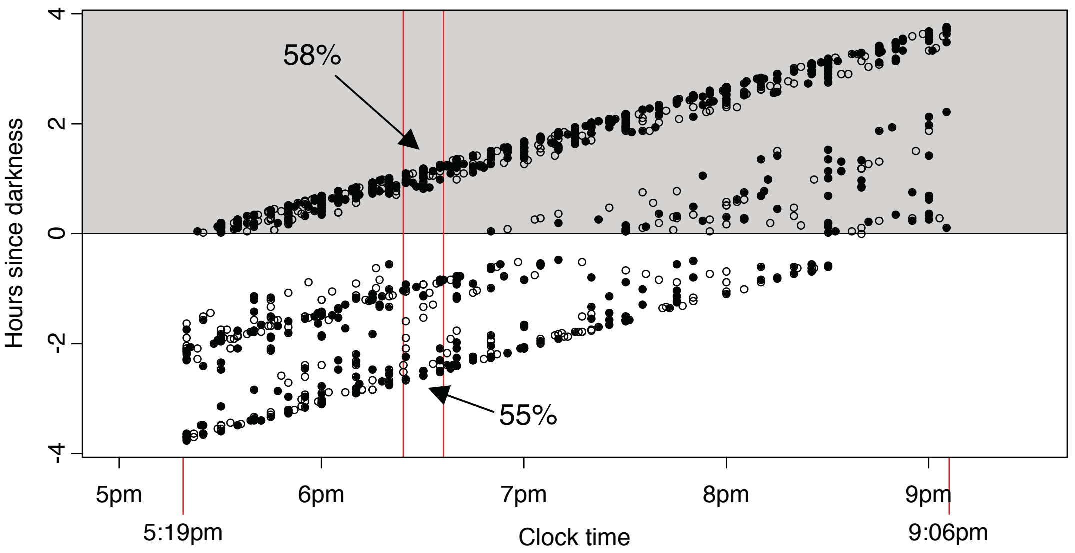


Figure 1: Plot of stops by clock time and darkness. The solid points indicate black drivers and the open circles represent non-black drivers. The shaded region indicates those stops occurring after the end of civil twilight. The large diagonal gap is a result of the shift from Pacific Daylight Time to Pacific Standard Time. The figure excludes stops occurring between sunset and the end of civil twilight. The vertical lines near 6:30 pm mark the example region discussed in the text. (Reproduced from Grogger and Ridgeway 2006)

The vertical lines in Figure 1 mark a period around 6:30 pm within which we can assess whether darkness influences the race of drivers stopped. During daylight hours 55% of the stops involved black drivers, while stops after dark involved black drivers in 58% of the stops, a slight difference and, if anything, runs counter to the racial profiling hypothesis.

Schell, Ridgeway and colleagues provide a similar analysis of three years of traffic stops in Cincinnati and find similar null conclusions against racial bias in traffic stop decisions.[xxxiii]

The instrumental variables approach here, however, does have limitations. First, this method assumes that the variation in daylight/darkness gives enough of a diminished capacity to effectively remove the importance of a suspect's race in the decision of whom to stop. If the police use car profiles, such as stylistic rims or other features that are correlated with race and social class, as the primary proxy for race then this approach will still yield an unbiased test of the race effect on police decisions but will be greatly underpowered because police will use these cues regardless of the level of daylight/darkness. Even if such proxies do not exist, the approach only measures the effect of race bias at those times of day that are sometimes light and sometimes dark. Since there is never daylight at 3am, we cannot estimate an effect of race for stops that occur at that hour.

INTERNAL BENCHMARKING

Recognizing the difficulty of assessing whether racial bias occurs on the aggregate in the decision to stop citizens has led some analysts to focus on the individual decision-making of police officers. The decision to stop a citizen is only one stage in the traffic stop process, at each stage at which police officers can introduce race bias in their decisions. Highly publicized examples of racial bias in police behavior can give an impression of systemic bias, even if the source of bias is only a few problem officers[xxxiv] (see Weitzer in this volume).[xxxv] The Christopher Commission in its assessment of abuse of police authority among the Los Angeles Police Department (LAPD), for example, noted that 10% of officers accounted for 27.5% of complaints of excessive force and 33% of all use of force incidents.[xxxvi]

[xxxiii] Schell, T., G. Ridgeway et al. Police-community relations in Cincinnati: Year three evaluation report.

[xxxiv] Jefferis, et al. The effect of a videotaped arrest on public perceptions of police use of force, pp. 381–395.

[xxxv] Weitzer, Incidents of police misconduct and public opinion, pp. 397–408.

[xxxvi] Christopher. Report of the Independent Commission on the Los Angeles Police Department.

The methods described previously which attempt to examine bias at the departmental level, are unlikely to detect the problem if the source is a small share of individual officers, and, even if somehow there are enough biased officers to create enough statistical power to detect the problem at the department level, these previous methods do not identify potential problem officers.

Walker[xxxvii] conceptualized the internal benchmark, a framework that compares officers' stop decisions with decisions made by other officers working in similar situational contexts. This method has been applied to department data in several localities and has been adopted as a part of several "early warning systems.[xxxviii]" At the Los Angeles Police Department (LAPD), the TEAMS II Risk Management Information System places officers in one of 33 peer groups.[xxxix] Officers in the same peer group presumably are expected to conduct similar policing activities. If an officer exceeds certain thresholds for their peer group, such being in the top 1 percent on number of complaints or number of use-of-force incidents, the system generates an "action item" for follow-up. However, officer roles in LAPD are certainly more diverse than 33 groups can capture. Similar problems are likely in other audit systems which compute a "peer-officer-based formula" to flag officers[xl] that does not take into account fully the variation in environments that officers in the same peer group work. Sometimes the peer group construction may be reasonable. For example, Decker and Rojek[xli] matched each St. Louis police officer to all other officers working in the same police districts. It is unclear whether matching by district alone was sufficient to ensure validity, although they argued that officers rotated shifts sufficiently so as not to warrant concern.

While this process is useful for flagging potential problem officers, it has some drawbacks. First, if officers in the entire precinct are equally biased, the method will not flag any officers as being problematic. We must rely on other analyses to assess that issue. Second, officers whom the method flags as outliers may have legitimate explanations for the observed differences. For example, a Spanish-speaking officer may appear to make an excessive number of stops of Hispanic

[xxxvii] See Walker, Searching for the denominator: Problems with police traffic stop data and an early warning system solution, pp. 63–95; Walker, S. The citizen's guide to interpreting traffic stop data: Un-raveling the racial profiling controversy. Unpublished manuscript; and Walker. Internal benchmarking for traffic stop data: An early intervention system approach.

[xxxviii] Walker. Early intervention systems for law enforcement agencies: A planning and management guide.

[xxxix] Birotte. Training evaluation and management system (TEAMS) II audit, phase I (fiscal year 2007/2008).

[xl] Walker. Early intervention systems for law enforcement agencies: A planning and management guide.

[xli] Decker, S., and J. Rojek, Saint Louis metropolitan police department traffic stop patterns.

suspects, when, in fact, the Spanish-speaking officer gets called in to handle and document those stops. Such situations should be detectable when supervisors review cases. Otherwise, the method eliminates possible explanations based on time or place, so the range of explanations is limited.

The fundamental goal of internal benchmarking is to compare the rate of nonwhite-pedestrian stops for a particular officer with the rate of nonwhite-pedestrian stops for other officers patrolling the same area at the same time. Matching in this way assures us that the target officer and the comparison officers are exposed to the same set of offenses and offenders.

Ridgeway and MacDonald[xlii] developed an internal benchmark methodology to compare the racial distribution of pedestrians/drivers whom individual police officers have stopped with that of pedestrians/drivers whom other officers in the same role have stopped at the same times and places. This method has been applied in case studies in both Cincinnati[xliii] and New York City.[xliv] Utilizing an approach based on propensity score weighting, doubly robust estimation, and false discovery rates these case studies attempt to customize the internal benchmark for each individual officer to a set of officers working in similar environments exposed to similar suspects and to control the risk too many officers being flagged as outliers (false positives). The first of the three stages in this process is, for each officer, to reweight the stops made by other officers so that they have the similar stop characteristics distributions.

Table 1 shows the results of this reweighting step for an example officer. Officer A made 392 stops. The method effectively identified 3,676 similarly situated stops made by other officers. These stops were selected as the benchmark group for Officer A because they were similar to Officer A's stops in terms of when they occurred (e.g., date, time of day), where they occurred (e.g., precinct, x-y coordinates), the assigned command of the officer making the stop, whether the officer making the stop was in uniform, and whether the stop was a result of a radio run. Figure 2 and Table 1 demonstrate that this collection of 3,676 is nearly identical to the officer's stops in several respects.

[xlii] Ridgeway, G. and J. MacDonald. Doubly robust internal benchmarking and false discovery rates for detecting racial bias in police stops.

[xliii] Ridgeway, G., et al. Police-community relations in Cincinnati: Year two evaluation report.

[xliv] Ridgeway, G. Analysis of racial disparities in the New York Police Department's stop, question, and frisk practices.

Table 1: Construction of an Internal Benchmark for a Sample Officer

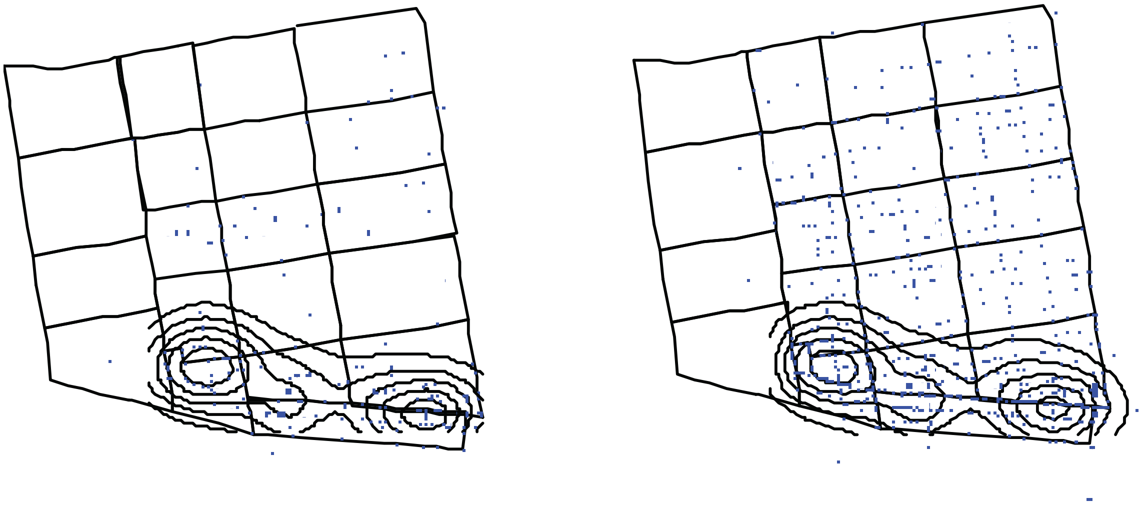
Stop Characteristic		Officer A (%)	Internal Benchmark (%)
		(N = 392)	(N = 3,676)
Month	January	3	3
	February	4	4
	March	8	9
	April	7	5
	May	12	12
	June	9	9
	July	7	7
	August	8	9
	September	10	10
	October	11	10
	November	11	11
	December	9	10
Day of the week	Monday	13	13
	Tuesday	11	10
	Wednesday	14	15
	Thursday	22	21
	Friday	15	16
	Saturday	10	11
	Sunday	15	14
Time of day	[12–2 a.m.]	11	11
	(2–4 a.m.]	5	5
	(10 a.m. –12 p.m.]	0	1
	(12–2 p.m.]	12	13
	(2–4 p.m.]	13	12
	(4–6 p.m.]	9	10
	(6–8 p.m.]	8	8
	(8–10 p.m.]	23	23
	(10 p.m. –12 a.m.]	17	17
Precinct	A	0	0
	B	98	98
	C	1	1
	D	1	0
Occurred inside?		4	6
Housing or transit	Transit	0	0
	Housing	0	0
	Other	100	100

Stop Characteristic		Officer A (%) (N = 392)	Internal Benchmark (%) (N = 3,676)
In uniform	Yes	99	97
Radio run	Yes	1	3

NOTE: The numbers in the table indicate the percentage of stops having that feature.

Furthermore, as shown in Figure 2, the distribution of the locations of the stops can be aligned geographically so that regions of this officer's stops in 2006 can be compared to other officers making stops in the same region.

Figure 2: Maps of the sample officer's stops and of similarly situated stops made by other officers



NOTE: The left map shows the sample officer's stops; the right map shows similarly situated stops made by other officers. The contours indicate the regions of the maps with the highest concentrations of stops.

An additional adjustment at this stage can improve the precision of this test. The second step of the process involves a regression model, to further refine the benchmark since some features are not perfectly matched between officers in Table 1, such as the frequency of being in uniform and being on a radio run.

Combining propensity score analysis with a second stage regression model has recently been labeled "doubly robust estimation" since if either the propensity score weights construct a well-matched set of benchmark stops or the regression model is correctly specified, then the resulting estimate of the officer's effect on the race of those stopped can be consistently estimated.[xlv]

The z-statistic from these regression models is the commonly used statistical measure for assessing the magnitude of the difference between an officer's minority stop fraction and the officer's internal benchmark group. The z-statistic scales the difference between the officer and his/her internal benchmark such that large differences based on a small number of stops are treated with greater uncertainty than large differences based on a large number of stops. Fridell[xlvi] suggests 2.0 and Smith[xlvii] suggests 1.645 as the appropriate z-scores to flag potentially problematic officers. However, such cutoffs generate too many false positives to be useful and are one of the sources of problems for LAPD's system. In a department of 1,000 officers we can expect 50 of them to have z-statistics in excess of 1.645 by chance alone.

Methods based on false discovery rates (fdr) helps address this kind of problem.[xlviii] The fdr is the probability of no difference between the officer and the benchmark given the value of an observed test statistic, z. We should flag those officers who have values of z that suggest a low probability of being incorrectly flagged as a problem. This approach when applied in Cincinnati noted 4 potentially problematic officers; and in New York City 15 potentially problematic officers.

Internal benchmark approaches provides a method for assessing individual officer bias. Again, the key to this approach is developing a reasonable peer group or comparison set of officers. This approach, however, is limited to departments with officers that make many stops. If officers make few stops (e.g. less than 50) then chance differences from their benchmark are likely and the comparisons are underpowered. Accumulating stops across years can improve this. For departments with few officers (e.g., those with less than 100 officers) the false discovery rate calculations become more unstable and more dependent on statistical assumptions.

[xlv] For technical details see Kang, J. and J. Schafer.Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data, pp., 523–580.

[xlvi] Fridell. By the numbers: A guide for analyzing race data from vehicle stops.

[xlvii] Smith, M. R. (2005). Depoliticizing racial profiling: Suggestions for the limited use and management of race in police decision-making, pp. 219–260.

[xlviii] Benjamini, Y. and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing.

POST-STOP OUTCOMES

The complexity of benchmarking for assessing bias in the decision to make a stop has in some cases caused analysts to abandon the endeavor in favor of assessing bias in post-stop outcomes, such as duration of the stop, decision to search, and use-of-force. This has its advantages since for this analysis we have a better assessment of the race distribution of who is at risk. However, substantial complexity remains.

Auditing police-citizen interactions

An obstacle to understanding racial disparities in police decision-making is that stopped drivers and pedestrians cannot observe how officers handle other stops, particularly those involving members of another race. They cannot answer the most pertinent question regarding racially bias policing, "would the same outcome have occurred if I had been a different race?" While such counterfactual questions so far have not been answered, recordings of stops can provide some guidance to understanding the dynamics in police-citizen interactions.

Dixon and colleagues[xlix] used a stratified random sample of 313 vehicle-mounted video and audio recordings from Cincinnati Police Department (CPD) cars to study interactions between police and community members. The study described how the race of the driver and the race of the officer influenced the dynamics of stops, including stop features associated with "counterproductive or dissatisfying interactions," and described how typical police–motorist interactions occur as a function of race.

Among the results reported in this study is the finding that interactions where the officer and driver are of the same race, officers are more likely to be interested in hearing the drivers' comments. The key problem that this creates in Cincinnati is that, since many more CPD officers are white, two-thirds of stops of black drivers involve a white officer while only one-third of stops of white drivers involve a black officer. Thus, the impact of degraded communication due to interracial stops will be greatest for the black drivers.

Additional research by the same research team[l] found that white officers conducted more investigative stops (e.g. asking questions about guns or drugs, asking for the IDs of passengers) while black officers were more likely to focus on the traffic infractions alone. Importantly, these differences did not depend on the

[xlix] Dixon, T. et al. The Influence of race in police–civilian interactions: A content analysis of videotaped interactions taken during Cincinnati police traffic stops, pp. 530-549.

[l] Schell, T. et al. Police-community relations in Cincinnati: Year three evaluation report.

race of the driver. That is, white officers also closely investigated white drivers. However, such differences between white and black officers can exacerbate the perception of racially biased policing. The black driver in Cincinnati who experiences one stop with a black officer and another stop with a white officer is likely to attribute the white officer's more intense investigation to race bias, even though on average this white officer treats blacks and whites with a similar level of scrutiny.

The analysis of recorded interactions is useful at identifying problem interactions, factors that can contribute to the perceptions of race, and stops that could be useful in training. However, such methods do not answer the question of whether the police use race as factor in deciding who to stop.

Hit rates

Hit rates, the percentage of conducted searches that turn up contraband, have been a frequently discussed outcomes test for racial equity in searches. If the hit rate for searched nonwhite suspects is less than the hit rate for searched white suspects, police might be applying a lower standard of suspicion to nonwhite suspects when deciding whether to search.

A series of papers by Persico and Todd[li] provide the theory and empirical examples of the use of hit rates with police traffic stop data. Relying on the premise of a Nash equilibrium, these authors argue that hit rates provide a race-neutral test of bias in police decision making because police decisions about which suspects to search take into account the benefits of searching different suspects, and suspects take "into account the risk of getting searched" (p. 37). [lii] If officers and criminals act as rational agents then selecting on the decision to stop someone the outcome of stops should be race neutral. Following on the logic of a Nash equilibrium that officers want to maximize their ability to find illegal contraband in traffic stops, and suspects want to reduce their likelihood of being caught, then the probability of successful "hits" should be equal once one conditions on the race of who is stopped. If, for example, police officers want to find illicit drugs and suspects want to avoid detection, the results for searches among police officers who are intentionally biased towards blacks will be offset by

[li] See Knowles, J., N. Persico, and P. Todd. Racial bias in motor vehicle searches, pp. 203–229; Persico, N., and P. Todd. Generalising the hit rates test for racial bias in law enforcement, with an application to vehicle searches in Wichita, pp. F351–F367; Persico, N., & Todd, P. (2006). Generalising the hit rates test for racial bias in law enforcement, with an application to vehicle searches in Wichita. The Economic Journal, 116, F351–F367; Persico, N. and P. Todd. The hit rates test for racial bias in motor-vehicle searches, pp. 37-53

[lii] Ibid.

a higher yield of searches among whites. In the long-run then the differences between races in hit rates should equalize. Perisco and Todd's analysis of Maryland State Police traffic stop data in several publications reports findings that the fraction of blacks stopped exceeds the fraction of black motorists on the road, but that the hit rates for the two groups is statistically equivalent.

We, however, provide an example to demonstrate that a simple comparison of hit rates can distort the true racial differences. Assume that suspects are stopped for either burglary or robbery. Further assume that there is no racial difference in the rates at which suspects carry contraband and that police are racially neutral in making stop and frisk decisions (essentially blind to race). Last, consider the information shown in Table 2. Within a crime category, hit rates are equal for black and white suspects. In this example, officers detain many more white suspects on suspicion of robbery, a crime with a higher hit rate, than they do black suspects, who are more likely to be stopped for burglary. In this example, though, those large differences in the rates of stops for burglary and robbery by race are due not to officer bias but are the result of racial differences in criminal participation. As a result, the total hit rate for white suspects is 4.6 percent ([1+45]/1,000), and, for black suspects, the hit rate is 1.4 percent ([9+5]/1,000).

Table 2: Hypothetical Example of a Hit-Rate Analysis

Race	Measure	Burglary	Robbery
White	Stopped and frisked	100	900
	Had contraband (%)	1	5
	Had contraband	1	45
Black	Stopped and frisked	900	100
	Had contraband (%)	1	5
	Had contraband	9	5

One could conclude from these two numbers (4.6% vs. 1.4%) that there is racial bias in the decision to search suspects, and that whites are not searched at sufficient rates. But officers in this hypothetical example are race neutral by design. Hit rates are equal across races for suspected burglars and equal across races for suspected robbers. This is a reminder that failing to account for an important factor—suspected crime, in this example—can distort the conclusions. In practice, the only way for the Nash equilibrium as described by Persico and Todd to work would be if black burglars and white robbers adjusted their criminal behaviors to mirror each other because they had equal probability of being stopped by the police.

This example illustrates a statistical problem that Ayres[liii] termed the subgroup validity problem, in which a particular relevant feature is more prevalent for certain racial groups. Other factors may impact the hit rate as well. Officers in some precincts may be likelier to frisk, due to crime in the area, recent surges in weapon recoveries, or a series of recent shootings, or more hostile attitudes displayed by suspects. An elevated frisk rate in some precincts may not meet with the community's approval, but it would be premature to attribute this variation to racial bias by police officers without examining other relevant factors. Therefore, it is critical to account for factors correlated with race that might be associated with both suspect race and the rate of contraband recovery.

In Ridgeway's analysis of hit rates in New York City, shown in Table 3, white and Hispanic suspects stopped in situations that were similar to the collection of black suspects had hit rates of 3.2 percent and 3.8 percent, respectively, compared with a hit rate of 3.3 percent for black suspects.[liv] There was no statistical evidence for a difference between these recovery rates. Furthermore, there were no differences in the rates at which officers found weapons on suspects. The unadjusted hit rates, however, suggested evidence of bias. Again, showing that it is important to adjust for subgroup differences in the circumstances by which different racial groups are subjected to police authority.

Table 3: Frisked or Searched Suspects Found Having Contraband or Weapons

	Black	Hispanic	White
Any contraband	3.3	3.2	3.8
Weapon	0.7	0.7	0.8

It is plausible that the carry rates, the percentage of stopped suspects that have contraband, differ by race. If white suspects simply carry drugs more frequently, perhaps believing that officers are unlikely to search them, then the contraband recovery rates for white suspects will be higher. Persico and Todd theorized from the logic of a Nash equilibrium that criminals will assess their risk of being searched and adjust their frequency of carrying drugs and weapons accordingly, so that an outcome test will be race neutral. It is difficult to confirm this in practice, and, as a result, conclusions drawn from Table 3 must allow for the possibility that carry rates are not uniform across racial groups.

[liii] Ayres. Outcome tests of racial disparities in police practices. pp. 131–142.

[liv] Ridgeway, G. Analysis of racial disparities in the New York Police Department's stop, question, and frisk practices.

Analysis of other stop outcomes

Other analysts have focused on developing appropriate benchmarks for studying the stop outcomes themselves. In Cincinnati, for example, Ridgeway[lv] notes that 47% of stops involving black drivers lasted less than 10 minutes while 56% of stops of nonblack drivers lasted less than 10 minutes. On the surface this seems to be a rather large bias. However, 18% of the stopped black drivers did not have valid drivers licenses while only 5% of nonblack drivers did not have valid licenses. As a result, we cannot discern whether the disparity in stop duration is attributable to the driver's race or to the additional time required to process a stop involving an unlicensed driver.

Social scientists recognize that adjusting for confounding variables is a critical step in all proper analyses, and there are clear examples in the current book where analysts attempt to make such adjustments (see Fagan and Davies; Parker and Stults in current volume). Particular to racial profiling analyses, police may approach vehicles more cautiously and conduct pat searches for weapons in high crime neighborhoods during peak crime times (e.g., late evening on the weekends). These decisions may occur regardless of the driver's race, but may be confounded with race due to differences in the neighborhoods by which minorities and whites live. In high crime neighborhoods police also may be more thorough in checking for vehicle registration and driver's license records, have a longer list of recent suspect descriptions that the stopped driver may match, and may be more likely to develop probable cause. In theory and practices all of these decisions could be independent of the driver's race. As a result, the stop location and time of the stop may influence all of the measured post-stop activities even in the absence of a race bias. When the race distribution of drivers differs by time and neighborhood location, one should adjust for these differences assessing racial bias in post-stop activity. The analysis also might adjust for other features occurring after the stop, such as whether the suspect had an open warrant or a suspended drivers license.

Location and time of the stop are two among a number of factors for which post-stop activity might vary that are confounded with race of drivers or pedestrians stopped by the police. While these differences may be structurally discriminatory based on racial differences in areas that individuals live, they may not be substantively discriminatory based on police-decision making.

The common practice of "adjusting for" potentially confounding factors with multivariate regression is difficult to defend in the analysis of post-stop data. The

[lv] Schell, T. et al. Police-community relations in Cincinnati: Year three evaluation report.

regression adjustment is only effectively if there is not a strong correlation between race and the other variables in the regression model. If in the case of citizen stops, the distribution of stop features of black differs substantially from the distribution of stop features of whites by neighborhood, type of violation, time of day, etc. it is uncertain whether the estimate of the race effect on police post-stop outcomes sufficiently accounts for these potentially confounding variables. Unless stops of black and white suspects occur in similar circumstances, the regression model will be sensitive to the terms in the model, such as interactions between race and other predictors (e.g., race*location). Unfortunately, this situation is often overlooked in criminological studies of racial profiling.

Earlier we showed an example in which we could reweight the stops of other officers to match the features of stops of a particular officer. In the same manner, Ridgeway (2006) showed that we can construct propensity score weights to reweight the stops of, for example, nonblack drivers or pedestrians to match the characteristics of the stops of black drivers or pedestrians. Table 4, from a Cincinnati Police Department study of racial profiling in traffic stops described in Schell, Ridgeway, and colleagues[lvi], provides a demonstration. The second column displays the percentages for the black drivers. The third column displays the percentages for the weighted non-black drivers.

Table 4: Comparison on a subset of stop features of the non-black driver sample to black drivers

	% Black drivers N = 20,146	% Non-black drivers (weighted) ESS = 5,365	% Non-black drivers (unweighted) N = 24,383
Neighborhood			
Downtown	2.4	2.4	4.8
Over-the-Rhine	7.1	6.9	3.2
I-71	2.1	2.1	6.1
I-75	6.0	6.1	13.6

[lvi] Ibid.

	% Black drivers N = 20,146	% Non-black drivers (weighted) ESS = 5,365	% Non-black drivers (unweighted) N = 24,383
Time of day			
12–3a.m.	23.3	21.8	16.7
3–6a.m.	5.2	4.8	3.7
6–9a.m.	6.0	8.3	10.8
9a.m.–12p.m.	6.8	7.8	12.7
12–3p.m.	6.9	7.5	12.8
3–6p.m.	16.9	17.8	15.2
6–9p.m.	15.8	14.9	12.7
9p.m.–12a.m.	19.0	17.0	15.4
Reason			
Equipment violation	24.0	22.6	12.7
Moving violation	66.1	69.7	83.4
Resident			
Cincinnati	91.8	90.8	63.2
Ohio (not Cincinnati)	3.8	4.3	18.8
Kentucky	1.9	2.6	11.7
Age			
Under 18	1.7	1.7	1.8
18-25	34.8	32.4	31.2
26-35	28.9	26.3	26.0
36-45	17.5	19.0	18.9
Invalid driver's license	18.0	13.2	5.3
Male	65.9	64.6	65.1

The weighted percentages for the non-black drivers are uniformly close to the percentages for the black drivers. Achieving this balance is the critical step when using propensity score techniques and removes the problems of insufficient overlap between races and non-linearity noted with regression models. Race, therefore, is the only factor differing between the groups by design. The fourth column in Table 4 displays the raw percentages for the non-black driver sample. These data indicate that very few non-black drivers are involved in stops in Over-the-Rhine.

Non-black drivers are much more likely to be stopped on the freeways. Therefore, the weighted sample has been constructed to downweight non-black drivers stopped on the freeways and upweight non-black drivers stopped in Over-the-Rhine. Additionally, non-black drivers with invalid drivers licenses are upweighted so that the rate of invalid drivers licenses in the comparison sample is closer to that of the black driver sample.

Aside from some statistical advantages, the method is also attractive for establishing the face validity of the method. Table 4 is easy to explain to a variety of policy audiences and it is effective for arguing that the subsequent results are based on apples-to-apples comparisons.

The raw numbers indicated that black drivers were much less likely than nonblack drivers to have had a traffic stop last less than 10 minutes, 47% versus 56%. After weighting, the nonwhite drivers stopped at similar times, places, and contexts had stops last less than 10 minutes 47% of the time, the same as the black drivers. All of the difference between the original numbers, 47% and 56%, can be attributable to the factors like time, place, and context.

There are advantages and disadvantages to both hit rates and matching approaches, like the propensity score approach previously discussed. The hit rate approach has intuitive appeal, providing a clear thought experiment where all else should be equal once the police make the decision whom to stop. The hit rates comparison assumes that selecting on who police decide to stop equalizes the two-groups so that whites and blacks should be equivalent. If blacks end up with lower hit rates than whites then one can argue the police are using a lower threshold in assessing suspicion for blacks. But is this reasonable? Actions transpire after the decision to stop that may be confounded with race. There is a body of research in criminology that suggests a variety of reasons for racial differences in stop outcomes. As we previously discussed, Dixon and colleagues found that black-white officer interactions in Cincinnati explained a substantial difference in the length of a stop and the decision to search a vehicle. These decisions, however, don't appear to be racially biased on the suspects but rather reflect racial differences in police officer practices. Engel and Tillyer[lvii] note the lengthy history of observation studies that find racial differences in suspect demeanor which too can effect outcomes in police-citizen interactions, such that all else but race is not equal once an officer has decided to stop a suspect.

[lvii] Engel, R.S. and R. Tillyer. Searching for equilibrium: The tenuous nature of the outcome test, pp. 54-71.

By contrast, matching approaches try to make all the statistical adjustments available with observational data. If one has the right set of variables then there is some confidence that a good test of the race-effect in post-stop outcomes can be assessed with accuracy. White and black suspects can be compared to each other in similar situations. If the analyst does not have the right set of contextual variables they can at least get better data and work on improving the matching strategy. There is no magic going on, no necessary thought experiment; one just wants to construct a feasible set of comparison groups.

CONCLUSIONS

The search for an appropriate method for assessing racial bias in police behavior has been a quest. Substantial improvements have been made as investigators have moved away from simple comparisons of police stop decisions to general populations estimates. The search for the appropriate benchmark, however, remains elusive. There is no clear way to establish the correct population at risk for police attention. All approaches have limitations. Clearly, the most feasible benchmarks are ones that attempt to remove as many factors that are potentially confounded with race as possible but are legally permissible on the part of the police. The key to drawing a causal inference about the importance of race is establishing a set of comparison conditions that are race neutral. This is, however, a significant challenge because many factors are highly confounded with race. Census estimates are inappropriate benchmarks. Observations are difficult to collect in a systematic fashion and require observers to note behaviors for which the police should consider someone suspicious. With enough training, effort, and time observation methods can be an effective benchmark in studies that focus on traffic enforcement on highways where minorities and whites are exposed to similar circumstances, but they are less likely to be useful in highly stratified urban environments where the police focus on much more than traffic enforcement. Arrest data is too confounded with police stop decisions to be a useful benchmark. After all, arrests are often a consequence of the decision to stop and search someone. Instrumental variables offer some promise by relying on variations in natures, such as the switch from daylight to darkness, that are independent of race. Here too instrumental variables are limited to drawing a causal inference from the conditions under which they are estimated. If, for example, the police behave systematically different towards minorities only in late night hours variations in natural daylight won't be useful for detecting racial bias. Hit rates are attractive because of the idea that police want to maximize their ability to find contraband and make reasonable arrests, so selecting on who is stopped should provide a race-neutral test. However, racial differences in the characteristics of criminal offenders can make a focus on hit rates invalid. Approaches that compare like criminals will yield better hit rate assessments. Matching approaches that compare whites to minorities in similar circumstances offer promise because they attempt to make apple-to-apple comparisons. A good matching approach, for example, could provide all relevant police factors net race. Omitted variables will always be a concern. What important variables are missing can, however, be a good subject of discussion. If the police cannot articulate a reasonable set of missing variables that are not recorded and are associated with racial differences in who is searched, the duration of stops, etc. then this provides at least circumstantial evidence of race bias.

Even if police decisions on whom to stop, search, detain are not intentionally biased they may be structurally discriminatory. Patrolling differently in high crime neighborhoods may place a disparate burden on minorities, but may not reflect actual bias in police decision-making, especially when one compares whites and minorities in similarly situated circumstances. Blacks, for example, disproportionately live in neighborhood plagued by crime and violence and there are few large US cities where whites live in comparable circumstances. Even when one does compare whites driving or walking through predominately minority neighborhoods and finds no difference in the probability of being stopped, searched, etc. the reality is that these individuals likely reflect only a small fraction of police actions in minority neighborhoods. So while the decisions by the police may not be intentionally biased, they may serve to affirm perceptions of bias because the level of police activity is greater in high crime-poverty areas disproportionately settled by minorities.

Unfortunately this is no unifying method that can establish the extent to which racially biased policing occurs. All approaches have weaknesses. Social scientists should therefore be measured in their assessments.

BIBLIOGRAPHY

Alpert, G. P., Smith, M. R., and Dunham, R. G. (2003), "Toward a Better Benchmark: Assessing the Utility of Not-at-Fault Traffic Crash Data in Racial Profiling Research," in Confronting Racial Profiling in the 21st Century: Implications for Racial Justice, Boston.

Alpert, Geoffrey, John M. MacDonald, and Roger Dunham. 2005. "Police Suspicion and Discretionary Decision Making During Citizen Stops." Criminology 43: 407-434.

Alpert, G. P., Smith, M. R., and Dunham, R. G. (2007), "Investigating Racial Profiling by the Miami-Dade Police Department: A Multimethod Approach," Criminology and Public Policy 6: 25-56.

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin (1996). "Identification of Causal Effects Using Instrumental Variables," Journal of the American Statistical Association 91(434):444-455.

Ayres, Ian (2002). "Outcome Tests of Racial Disparities in Police Practices," *Justice Research and Policy*, 4, pp. 131–142.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, Series B 57, 289–300.

Birotte, A. (2007, November). Training evaluation and management system (TEAMS) II audit, phase I (fiscal year 2007/2008). Technical report, Office of the Inspector General, Los Angeles Police Department, Los Angeles, CA. http://www.lacity.org/oig/Reports/TEAMS2F1Report_11-06-07.pdf.

Christopher, Warren. (1991). Report of the Independent Commission on the Los Angeles Police Department.

Decker, S., and Rojek, J. (2002). *Saint Louis metropolitan police department traffic stop patterns*. Report submitted to the St. Louis Police Department, January.

Dixon, Travis L., Terry L. Schell, Howard Giles, and Kristin L. Drogos (2008). "The Influence of Race in Police–Civilian Interactions: A Content Analysis of Videotaped Interactions Taken During Cincinnati Police Traffic Stops," Journal of Communication 58:530-549.

Eck, John E., Lin Liu, and Lisa Growette Bostaph, *Police Vehicle Stops in Cincinnati: July 1–December 31, 2001*, 2003. Online at http://www.cincinnati-oh.gov/police/downloads/police_pdf6937.pdf as of November 6, 2005.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. Journal of the American Statistical Association 99, 96–104.

Efron, B. (2007). Correlation and large-scale simultaneous significance testing. Journal of the American Statistical Association 102 (477), 93–103.

Engel, Robin S. and Tillyer, Rob (2008) 'Searching for Equilibrium: The Tenuous Nature of the Outcome Test', Justice Quarterly, 25:1, 54—71.

Farrell, Amy, Dean Jack McDevitt, Shea Cronin, and Erica Pierce (2003). Rhode Island Traffic Stop Statistics Act: Final Report. Available at http://www.racialprofilinganalysis.neu.edu/IRJ_docs/RIFinalReport.pdf

Fridell, Lorie A., By the Numbers: A Guide for Analyzing Race Data from Vehicle Stops, Washington, D.C.: Police Executive Research Forum, 2004. As of November 9, 2007: http://www.policeforum.org/upload/BytheNumbers%5B1%5D_715866088_12302005121341.pdf

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

Greenwald, H. P. (2001), "Final Report: Police Vehicle Stops in Sacramento, California," Sacramento Police Department.

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, 878-887.

Imbens, G. (2003). "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: a Review," Technical Working Paper 294, National Bureau of Economic Research.

Hindelang, Michael. "Variations in Sex-Age-Race Incidence Rates of Offending." American Sociological Review 46: 461-475.

Jefferis, Eric S., Robert J. Kaminski, Stephen Holmes, and Dena E. Hanley, "The Effect of a Videotaped Arrest on Public Perceptions of Police Use of Force," Journal of Criminal Justice, Vol. 25, No. 5, 1997, pp. 381–395.

Kang, J. and J. Schafer (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science* 22 (4), 523–580.

Klinger, David A., "Demeanor or Crime? Why 'Hostile' Citizens Are More Likely to Be Arrested,"Criminology, Vol. 32, No. 3, 1994, pp. 475–494.

———, "Negotiating Order in Patrol Work: An Ecological Theory of Police Response to Deviance,"Criminology, Vol. 35, No. 2, 1997, pp. 277–306.

Klinger, David A. and George Bridges. 1997. "Measurement Error in Calls-For-Service as an Indicator of Crime." Criminology, 35(4): 705-726.

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

Lamberth, J. (1994), "Revised Statistical Analysis of the Incidence of Police Stops and Arrests of Black Drivers/Travelers on the New Jersey Turnpike Between Exits or Interchanges 1 and 3 From the Years 1988 Through 1991," report, Temple University, Dept. of Psychology.

Lamberth, John (2003). "Racial Profiling Data Analysis Study: Final Report for the San Antonio Police Department," available at http://www.sanantonio.gov/sapd/pdf/LamberthSanAntonioRpt_2003.pdf.

Lange, J. E., Blackman, K. O., and Johnson, M. B. (2001). Speed violation survey of the New Jersey turnpike: Final report. Office of the Attorney General of New Jersey.

Maxfield, Michael and George L. Kelling (2005). New Jersey State Police and Stop Data: What Do We Know, What Should We Know, and What Should We Do? The Police Institute at Rutgers-Newark, School of Criminal Justice, Rutgers University.

McConnell, E. H., and Scheidegger, A. R. (2001), "Race and Speeding Citations: Comparing Speeding Citations Issued by Air Traffic Officers With Those Issued by Ground Traffic Officers," paper presented at the annual meeting of the Academy of Criminal Justice Sciences, Washington, DC, April 4–8.

Montgomery County Department of Police (2002), "Traffic Stop Data Collection Analysis," 3rd report.

Neyman, Jerzy (1923). On the application of probability theory to agricultural experiments: Essay on principles, Section 9," translated in Statistical Science 5, 465-480, 1990.

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

Persico, Nicola and Todd, Petra E. (2008) "The Hit Rates Test
for Racial Bias in Motor-Vehicle Searches", Justice Quarterly, 25:1, 37 — 53.

Reiss, Albert J. (1971). Systematic observation of natural social phenomena. Sociological Methodology, 3: 3-33.

Ridgeway, G. (2006). Assessing the effect of race bias in post-traffic stop outcomes using propensity scores. Journal of Quantitative Criminology 22 (1),1–26.

Ridgeway, G., T. L. Schell, K. J. Riley, S. Turner, and T. L. Dixon (2006). Police-community relations in Cincinnati: Year two evaluation report. Technical Report TR-445-CC, RAND Corporation, Santa Monica, CA. http://www.rand.org/pubs/technical_reports/TR445/.

Ridgeway, G. (2007). Analysis of racial disparities in the New York Police Department's stop, question, and frisk practices. Technical Report TR-534-NYCPF, RAND Corporation.

Ridgeway, G. and J. M. MacDonald (2008). Doubly robust internal benchmarking and false discovery rates for detecting racial bias in police stops. Working paper.

Sampson, Robert J., and William Julius Wilson, "Toward a Theory of Race, Crime, and Urban Inequality," in John Hagan and Ruth D. Peterson, eds., Crime and Inequality, Stanford, Calif.: Stanford University Press, 1995, pp. 37–54.

Schell, Terry L., Greg Ridgeway, Travis L. Dixon, Susan Turner, K. Jack Riley (2007). Police-Community Relations in Cincinnati: Year Three Evaluation Report. RAND TR-535-CC. http://www.rand.org/pubs/technical_reports/TR535/

Skogan, Wesley G., Disorder and Decline: Crime and the Spiral of Decay in American Neighborhoods, Berkeley, Calif.: University of California Press, 1990.

Smith, Douglas A., "The Neighborhood Context of Police Behavior," in Albert J. Reiss and Michael H. Tonry, eds., Communities and Crime, Chicago: University of Chicago Press, 1986, pp. 313–341.

Smith, M. R. (2005). Depoliticizing racial profiling: Suggestions for the limited use and management of race in police decision-making. *George Mason University Civil Rights Law Journal* 15 (2), 219–260.

Walker, S. (2001). Searching for the denominator: Problems with police traffic stop data and an early warning system solution. *Justice Research and Policy* 3 (2), 63–95.

Walker, S. (2002). The citizen's guide to interpreting traffic stop data: Un-raveling the racial profiling controversy. Unpublished manuscript.

Walker, S. (2003a). Early intervention systems for law enforcement agencies: A planning and management guide. Technical report, Office of Community Oriented Policing Services, U.S. Department of Justice, Washington DC. http://www.cops.usdoj.gov/html/cd_rom/inaction1/pubs/EarlyInterventionSystemsLawEnforcement.pdf.

Walker, S. (2003b). Internal benchmarking for traffic stop data: An early intervention system approach. Technical report, Police Executive Research Forum.

Weitzer, Ronald, "Incidents of Police Misconduct and Public Opinion," *Journal of Criminal Justice*, Vol. 30, No. 5, 2002, pp. 397–408.

Zingraff, M. T., Mason, M., Smith, W., Tomaskovic-Devey, D., Warrent, P., McMurray, H. L., and Fenlon, R. C. (2000), "Evaluating North Carolina State Highway Patrol Data: Citation, Warnings, and Searches in 1998," report submitted to North Carolina Department of Crime Control and Public Safety and North Carolina State Highway Patrol.