Available online at www.sciencedirect.com
ScienceDirect
Journal of Criminal Justice 36 (2008) 138–153
JOURNAL OF CRIMINAL JUSTICE

The intersection of racial profiling research and the law

Rob Tillyer *, Robin S. Engel, John Wooldredge

Division of Criminal Justice, University of Cincinnati, P. O. Box 210389, Cincinnati, OH 45221, United States

Abstract

There has been a significant increase in the litigation of selective enforcement cases based on racial profiling claims. This trend has resulted in two legal issues that are problematic for racial profiling research. First, selective enforcement claims that rely on statistical evidence must successfully measure "similarly situated persons" who were eligible for police stops to provide a comparison against those actually stopped by police. Second, the research must demonstrate "how much" statistical evidence of racial/ethnic disparities exists. Although these legal components are necessary for successful selective enforcement claims, the methodologies and statistical analyses currently used in racial profiling research cannot adequately address these issues. It is argued that the over-reliance on social science research, in general, and statistical techniques, specifically, to provide evidence of discrimination in selective enforcement cases places policing research and legal decision making at a crossroads.

Introduction

The intersection of policing research and the law is not a new phenomenon. Previous research on police use of force and police response to domestic violence situations resulted in both legal challenges and changes in formal organizational policies (e.g., Fyfe, 1982, 1988; Lempert, 1989; Sherman & Berk, 1984; Sherman & Cohen, 1989). Yet, the recent emphasis on racial profiling, in combination with litigation regarding claims of biased police behavior, has moved social science research even closer to the legal arena. M. R. Smith and Alpert (2002) outlined the difficulties in researching police stopping behavior, and suggested that the use of this research can provide judges with increased evidence to make decisions regarding claims of selective enforcement based on race/ethnicity. Additional years of data collection and analyses suggest, however, that the limitations of social science research are significant and the vigor of its application in the legal arena should be tempered.

While the use of social science research can assist in legal decisions, it must be used with caution due to an inability to measure the reasons for officer decision-making. Although social science research and statistical analyses of police behavior are useful for identifying racial and ethnic *disparities* in traffic and pedestrian stops, their reliability for determining racial and ethnic *discrimination* is unknown.[1] The differences in these two concepts is based on the individual intentions of police officers—in one case, racial/ethnic discrimination is the result of factors other than individual police bias, while in the other, racial/ethnic discrimination is the direct result of intentional police bias. This is especially problematic given that current social scientific research examining "racial profiling" was developed as a response to selective enforcement legal claims, which often include the use of statistics to demonstrate discriminatory purpose and effect.

The problematic use of social science research in selective enforcement litigation hinges on two necessary legal components that have not yet been successfully broached with the current methodologies employed in racial profiling research. First, selective enforcement claims that rely on statistical evidence must successfully measure "similarly situated persons" who were eligible for police stops to provide a comparison against those actually stopped by police. Second, the research must demonstrate "how much" statistical evidence of racial/ethnic disparities exists. Although these legal components are necessary for successful selective enforcement claims, the methodologies and statistical analyses currently used in racial profiling research do not adequately address these issues. The result is a body of research that can inform the legal system, police, politicians, and citizens regarding patterns or trends of

* Corresponding author. Tel.: +1 513 556 0615; fax: +1 513 556 2037.
E-mail address: rob.tillyer@uc.edu (R. Tillyer).

0047-2352/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jcrimjus.2008.02.004

potential racial/ethnic *disparities* by police, but perhaps should not be used to determine racial/ethnic *discrimination* by police.

These issues are further examined in this article, which begins with a brief review of the history of racial profiling, including its original goals and subsequent transformation into an area of significant concern for policymakers, politicians, police departments, and social science researchers. Thereafter, the legal decisions that have shaped the intersection between racial profiling research and the law are reviewed, including the relevance of using statistics and the amount of evidence necessary to show discriminatory purpose and/or effect. Flowing from this discussion, this article then highlights the validity and reliability of various methods for creating benchmarks against which to measure police officers' stopping behavior. Finally, the amount of evidence necessary to determine discriminatory purpose and effect is discussed, and the two-step process typically used for statistically identifying racial and ethnic disparities in selective enforcement cases is critiqued. The article concludes by highlighting the problems associated with the over-reliance on social science research, in general, and statistical techniques, specifically, to provide evidence of discrimination in selective enforcement cases and provides several suggestions for how social science might improve the study of racial profiling. Collectively, this article provides a review and critique of common social science techniques and their use in studying traffic stops and informing legal decisions regarding disparity and/or discrimination.

History of racial profiling research

In regard to understanding police discretion, the study of police-citizen interactions has a research history of over fifty years. Initially, the 1957–58 American Bar Foundation Survey of criminal justice agencies highlighted the extensive use of discretion within the criminal justice system, and in particular by police officers (Walker, 1993). The result of this in-depth inquiry into the criminal justice system was the identification of potential extralegal factors that might influence criminal justice actors' behaviors. Reaction to this series of studies initiated a body of research that emerged to understand more thoroughly why and how police officers use their discretion. While early studies were often qualitative in nature (Bittner, 1970; Rubinstein, 1973; Skolnick, 1966; Van Maanen, 1974), the research has become heavily quantitative over the past forty years (for review see National Research Council, 2004; Riksheim & Chermak, 1993).

More recent quantitative studies have indicated a stronger influence of legal variables on police-citizen encounters compared to other nonlegal factors (e.g., Alpert & Dunham, 2004; Engel & Silver, 2001; Klinger, 1994; Mastrofski, Worden, & Snipes, 1995; National Research Council, 2004; Novak, Frank, Smith, & Engel, 2002). Although some studies have found citizens' race/ethnicity to be a significant factor in coercive police action (e.g., citations, arrests, and use of force), the majority of studies that reported race effects also noted these effects were relatively small and account for little of the explained variance in the statistical models (e.g., Engel & Silver, 2001; Novak et al., 2002; Riksheim & Chermak, 1993; Terrill & Mastrofski, 2002; Worden, 1989; for review, see National Research Council, 2004).

In contrast to the larger body of research on police decision-making, "racial profiling" research was prompted by high-profile litigation, political pressure, widespread public disapproval of policing tactics, and recommendations from social scientists. Unfortunately, the result is a body of research that generally fails to ask the proper questions, tends to be methodologically weak, and has been used inappropriately by social scientists, the media, politicians, and the courts (Engel, Calnon, & Bernard, 2002; M. R. Smith & Alpert, 2002). These shortcomings are evident throughout the history of racial profiling research in the United States.

The practice of targeting racial minorities for routine traffic and pedestrian stops originated with the war on drugs and promoted profiling as an effective policing tactic to detect drug offenders (Harris, 2002; Tonry, 1995). The first racial profile originated from the attempt to interdict the flow of drugs from Miami up Interstate 95 to the cities of the Northeast (Harris, 1999). As noted by Harris, a report produced by the Drug Enforcement Administration (DEA) concluded that "large scale, interstate trafficking networks controlled by Jamaicans, Haitians, and Black street gangs dominate the manufacture and distribution of crack." In 1986, the DEA established "Operation Pipeline," a highway drug interdiction program designed to train federal, state and local law enforcement officials on the indicators of drug trafficking activities of motorists (General Accounting Office, 2000). One of the indicators of drug trafficking used in the training was the race/ethnicity of the driver. Some academics and group activists argued that the use of these and other training materials generated a racially and ethnically biased drug courier profile and encouraged the targeting of minority motorists for traffic stops (e.g., American Civil Liberties Union, 1999; Harris, 2002).

The perceived legitimacy of these law enforcement tactics, however, was short-lived. In the 1990s, the alleged use of racial profiles by the Maryland State Police and the New Jersey State Police resulted in high profile and successful litigation by those claiming constitutional violations based on selective enforcement (e.g., *State of New Jersey v. Soto et al.*, 1996; *Wilkins v. Maryland State Police et al.*, 1993). The combined effects of successful litigation, pressure from politicians and public interest groups, and widespread media attention surrounding the issue of racial profiling led to a crisis of legitimacy for police departments across the country (Engel & Calnon, 2004b). In an effort to prove or disprove the practice of racial profiling, the courts responded by requiring the collection of data in police departments facing litigation. These data collection efforts were supported by the federal government and the social science community, who argued that collecting information about the characteristics of the citizen during police-citizen encounters would enable researchers to determine if police officers were engaging in racial profiling (General Accounting Office, 2000; Ramirez, McDevitt, & Farrell, 2000).

Police departments responded to this crisis of legitimacy in three ways. First, police departments across the country issued policy statements and formal orders prohibiting the use of race/ethnicity characteristics as criteria for police decision-making. Second, they included or increased the amount of racial sensitivity training in their academy and in-service training curricula. Third,

and most importantly, departments began collecting data on traffic and pedestrian stops. In some cases, these responses from police administrators were court ordered or legislatively mandated, while in other agencies, the changes were made voluntarily. Importantly, however, the widespread collection of traffic stop data was *not* developed to understand police behavior, but to address police administrators' fears of litigation, political pressure, local and state statutes, and court orders. As a result, the methodologies of racial profiling data collection efforts developed almost entirely around issues related to selective enforcement litigation. That is, the social scientific community designed methodologies specifically to provide evidence regarding discrimination for the legal community, rather than to understand police behavior.

M. R. Smith and Alpert (2002) documented several ways in which social scientists could aid the collection of information for selective enforcement litigation. Nevertheless, there are two legal issues that social scientists have been unable to successfully navigate. First, measuring similarly situated persons; and second, determining the amount of statistical evidence necessary to demonstrate discriminatory purpose and effect. These two legal issues, and their corresponding methodological and statistical issues, are detailed below.

Legal considerations in racial profiling research

The court held in *Whren et al. v. U.S.* (1996) that pretextual stops were not a violation of the Fourth Amendment protection against unreasonable searches and seizures. Thus, for purposes of the Fourth Amendment, whether or not racial animus was involved in officers' decisions to make traffic stops is irrelevant (M. R. Smith & Alpert, 2002). As a result of the court's decision in *Whren*, current racial profiling litigation is often based on claims of selective enforcement, and is generally litigated as equal protection cases under the Fifth and Fourteenth Amendments. The typical remedies sought in selective enforcement equal protection cases are suppression of evidence in criminal cases, or damages in civil suits brought under 42 U.S.C. Section 1983 (M. R. Smith & Alpert, 2002). To be successful in claiming selective enforcement, the plaintiff needs to show "discriminatory purpose" and "discriminatory effect" (*McCleskey v. Kemp*, 1987; *U.S. v. Armstrong*, 1996). In *McCleskey v. Kemp* (1987), the court detailed the challenges facing a plaintiff claiming a constitutional violation under the equal protection clause including the difficulty of proving discriminatory purpose with statistics. Regardless of the difficulties, the use of statistics to show a discriminatory purpose and effect for similarly situated persons, and determining the amount of evidence necessary to show discrimination are the two salient issues for racial profiling research. These two legal issues frame the methodological issues involved in racial profiling research and reflect the problematic intersection of social science and the law.

Use of statistics to show discriminatory purpose and/or effect

Selective enforcement cases require evidence of discriminatory purpose and discriminatory effect. Discriminatory purpose is difficult to prove, as the most direct means by which it is established is through an officer's admission that he/she intentionally discriminated against the citizen (a highly unlikely event). As a result, statistics are a more realistic avenue to support a claim of discriminatory purpose. Similarly, evidence for discriminatory effect is likely to be accessed through statistical analyses. Comparisons of "persons in similar circumstances," also referred to as "similarly situated persons," has become an important component for establishing the discriminatory effect based on statistical comparisons (see *Wo v. Hopkins,* 1886). That is, statistical comparisons are made between those individuals who *had* the law enforced upon them versus those individuals who *could have* had the law enforced upon them but did not (i.e., those eligible, but not selected for enforcement).

Historically, the use of "similarly situated individuals" analyses occurred in jury selection cases, and these cases have routinely used statistical evidence to show discriminatory effect. For example, in *Turner et al. v. Fouche et al.* (1970), the Supreme Court unanimously agreed that a statistical comparison of 60 percent Black county population (those eligible to be selected for the grand jury) to 37 percent Blacks placed on the grand jury list was evidence of racial discrimination. Specifically, the Court held that the appellants demonstrated a "substantial disparity between the percentage of Negro residents in the county as a whole and of Negros on the newly constructed jury list" (*Turner et al. v. Fouche et al.*, 1970, p. 360). In the *Turner* case, persons in similar circumstances were easily measured. That is, it was relatively straightforward to identify those people who were eligible for selection to the jury list but who were not selected. To be on the grand jury list, individuals had to be county residents and could have been eliminated from the list by jury commissioners if they were not "upright" and "intelligent." The appellants argued that the county census data was a reliable source for comparison data (to demonstrate similarly situated persons) given that the only county residents not eligible for selection were those considered to be not "upright" or "intelligent."[2]

Eight years later, in a five to four decision, the U.S. Supreme Court further extended the use of statistics to show discrimination (*Castaneda v. Partida*, 1977). As in *Turner*, the Court in *Castaneda* allowed the use of population statistics to establish a prima facie case of discrimination in the grand jury selection process in Texas. In this case, 79 percent of the population in the county was Mexican-American compared to 39 percent of those selected for grand jury service over an eleven-year period. The statute in this case required that to be eligible for grand jury service, individuals had to be citizens of the county, qualified voters in the county, of sound mind, good moral character, literate, with no prior felonies, and no pending indictments or other accusations of theft or any felony. The comparison to "similarly situated persons" was made through the use of census data that eliminated foreign born, and persons over twenty-five who had no schooling (for the literacy component of the statute). The eligible population of Mexican-Americans decreased from 79 percent to 65 percent based on these criteria, however, the Court found this still represented a significant disparity compared to grand juries comprised of 39 percent Mexican-Americans.

Importantly, in *Castaneda*, the ability to validly measure "similarly situated persons" was somewhat reduced. In cases where the only requirement for jury selection was county residency, population figures represented an exact measure of "similarly situated persons" who were eligible for jury selection. As statutes require additional criteria for jury selection, however, social scientists' ability to provide an exact measure of similarly situated persons is diminished. The holding in *Castaneda* represented a shift in the strict use of "known" populations as comparisons groups. As noted by Chief Justice Burger in his dissent "the decisions of this Court suggest, and common sense demands, that the *eligible* population statistics, not gross population figures, provide the relevant starting point" (*Castaneda v. Partida*, 1977, p. 504, emphasis added).

The need to measure similarly situated persons is a critical issue for using statistical comparisons to demonstrate a discriminatory purpose and effect in selective enforcement cases. As described in detail below, measuring similarly situated persons for the purposes of demonstrating discrimination in racial profiling cases is more difficult than simply making comparisons to residential population statistics, which is the technique routinely used in jury selection cases. Measuring the eligible population is more difficult in racial profiling research due to the difficulty in determining an accurate benchmark against which police behavior can be compared.

Amount of evidence necessary to show discriminatory purpose and/or effect

Once it has been established that some statistical evidence exists that demonstrates discriminatory purpose and/or effect, the question then becomes, how much evidence is necessary; that is, how strong must the statistical findings be to demonstrate discrimination? The U.S. Supreme Court has generally used the terms "some evidence," "clear evidence," and "credible showing" to describe the strength of the evidence necessary to establish a prima facie case for discrimination (*U.S. v. Armstrong,* 1996). The amount of evidence necessary is decided on a case-by-case basis, taking into account the totality of the evidence presented. As stated in *Alexander v. Louisiana* (1972), "this Court has never announced mathematical standards for the demonstration of 'systematic' exclusion of blacks, but has rather, emphasized that a factual inquiry is necessary in each case that takes into account all possible explanatory factors" (p. 630). Of course, the obvious problem with this approach is the lack of consistency across cases when determining whether or not there is ample statistical evidence to demonstrate discriminatory effect.

The U.S. Supreme Court has accepted the use of summary statistics—including standard deviations—to make determinations regarding the strength of the evidence. For example, in *Castaneda v. Partida* (1977), the court held that "as a general rule for such large samples, if the difference between the expected value and the observed number is greater than two or three standard deviations, then the hypothesis that the jury drawing was random would be suspect to a social scientist" (p. 496). As will be demonstrated below, however, the use of a standard deviation to demonstrate discriminatory effect in selective enforcement cases based on traffic and pedestrian stops is likely inappropriate.

To summarize, there are two issues in selective enforcement litigation—measuring similarly situated persons and determining the amount of evidence necessary to establish discriminatory effect—that have had a profound impact on social science research. Prior to the traffic stop studies that originated in the 1990s, the bulk of policing research was driven by the social science inquiry of explaining police decision-making. In contrast, the aforementioned two legal issues have driven the methodologies and development of racial profiling research, not for greater understanding of police behavior, but to establish evidence of discrimination for use in litigation. The following sections detail several problems that arise when statistical procedures are utilized to achieve this goal.

Measuring similarly situated persons in racial profiling research: problems with benchmarking

A major limitation of traffic stop data analyses is the inability to make firm conclusions regarding discrimination. One reason for the difficulty in making these conclusions is that the driving population eligible to be stopped by police is unknown. Determining how often minorities are stopped by police is not particularly meaningful until those percentages are compared to some "expected probability" of these actions toward minorities (Fridell, 2004; Rojek, Rosenfeld, & Decker, 2004; M. R. Smith & Alpert, 2002). These expected probabilities are often referred to by academics as "benchmarks," "base rates," "baselines," or "denominators." In the legal arena, the benchmark represents a comparison to "similarly situated persons" as required for using statistics to demonstrate discriminatory effect in selective enforcement cases (see *Castaneda v. Partida,* 1977; *Turner et al. v. Fouche et al.,* 1970; *U.S. v. Armstrong,* 1996; *Wo v. Hopkins,* 1886). Studies examining racial disparities compared police stop data with the "expected" rate of stops of minorities assuming that no racial discrimination or prejudice exists by police. That is, these studies attempted to compare drivers who were actually stopped by police to drivers who were *eligible* to be stopped by police.

The expected rate of minority stops has been estimated through several different benchmarks, including residential census populations, "adjusted" census populations, statistically derived traffic flow models, not-at-fault traffic accidents, citizen surveys, internal comparisons to other officers, observations of traffic, and observations of traffic law-violating behavior (Engel & Calnon, 2004a; Fridell, 2004). Specifically, residential census data provide the racial composition of the population living in that jurisdiction and are used as a proxy for the driving population. Second, the "adjusted" census populations are created from the census population figures, but are weighted in an attempt to reflect that individuals drive in areas other than where they reside. In other words, based on the census data, the racial composition of an area is modified to reflect the limitation that the raw census population estimates do not consider drivers from other jurisdictions (see Farrell, McDevitt, Bailey, Andresen, & Pierce, 2004; Farrell, McDevitt, Cronin, & Pierce, 2003; Novak, 2004; Rojek et al., 2004). A third type of benchmark is the use of traffic flow models.

These are similar to the "adjusted" census population in that they use the residential census population data; in addition, they also incorporate other information, such as race information gathered from traffic stops, to develop an estimate of the driving population (see Eck, Liu, & Bostaph, 2003; Engel et al., 2005). Fourth, not-at-fault accidents can be used as a proxy for driving behavior as the distribution of these accidents is argued to randomly occur, thereby providing a measure of the racial composition of the drivers (see Alpert, Smith, & Dunham, 2004; Alpert Group, 2004). Fifth, citizen surveys are used to directly access information from individuals and can take three general forms: (1) survey inquiries regarding the citizen's interaction with the police (e.g., Langan, Greenfeld, Smith, Durose, & Levin, 2001), (2) survey inquiries regarding individuals' driving patterns (e.g., Boyle, Dienstfrey, & Sothoron, 1998; Pickrell & Schimek, 1998), and (3) surveys of drivers as they exit roadways (e.g., Lange & Voas, 2000). A sixth method compares the behavior of officers or groups of officers to other similarly situated officers to assess if there is internal consistency among the officers in stopping behavior (e.g., Walker, 2001). This method does not attempt to measure similarly situated persons.

The final two methods, observations of driving behavior and observations of traffic law-violating behavior, involve placing trained observers on the highway to physically count the racial composition of drivers traveling on selected roadways (e.g., Engel, Calnon, Liu, & Johnson, 2004; Lamberth, 2003). These methods differ only because the latter includes measures of traffic violations in addition to roadway usage (e.g., Engel et al., 2004; Lamberth, 1994, 1996; W. R. Smith et al., 2003). Collectively, it is clear that some of these benchmarking techniques are more valid indicators of the expected rate of minority stops than others.[3] Unfortunately, as demonstrated below, none of the benchmarking techniques have produced reliable and consistent data that adequately measure the expected rate of stopping specific racial groups.

Inconsistencies across benchmarks

One of the most problematic issues regarding the use of traffic stop statistics as evidence of discriminatory effect in selective enforcement cases is the inconsistency across benchmark comparisons. As noted above, several different benchmarking techniques are readily used to establish estimates of the driving population at risk for motor vehicle stops to compare with actual traffic stops. Depending on the benchmark selected, the findings of racial/ethnic disparity can vary dramatically. The following example using data collected for the Project on Police-Citizen Contacts commissioned by the Pennsylvania State Police vividly demonstrates the tremendous variation across benchmark comparisons.

The Project on Police-Citizen Contacts involved the collection of data by Pennsylvania State Troopers on all member-initiated traffic stops, and comparison of these data to multiple benchmarks to determine the amount of racial/ethnic disparity in traffic stops across counties (for details regarding the methodology and findings of this study, see Engel & Calnon, 2004a; Engel et al., 2004; Engel et al., 2005). To examine racial/ethnic disparities in member-initiated traffic stops, five different comparisons were made: (1) all traffic stops were compared to the residential census population, (2) traffic stops of only drivers who resided in the counties where the stop occurred were compared to the residential census population, (3) all traffic stops were compared to traffic flow model estimates derived from residential populations, (4) daytime traffic stops were compared to daytime traffic observations, and (5) daytime speeding traffic stops were compared to daytime speeding observations.[4] Disproportionality ratios that measure the odds of stopping a minority in comparison to the odds of stopping a White driver were created based on these comparisons for twenty-seven of the sixty-seven counties in the Commonwealth of Pennsylvania.[5] The findings reported in Table 1 showed dramatic range in the value of the disproportionality ratios within the same counties.

One example dramatically illustrates the inconsistencies across benchmarks. In Jefferson County, Pennsylvania, the findings illustrated that the same traffic stop data compared to different benchmarks resulted in interpretations that ranged from Black motorists reported to be 69.2 times *more* likely than Caucasians to be stopped for traffic offenses (based on census data), compared to Black motorists actually being *less* likely than Caucasians to be stopped (based on the traffic flow model, 0.3). In other words, the range between disproportionality ratios within the same county varied from 69.2 for the census-based model to 0.3 for the traffic flow model. This county is characterized by small minority residential populations (0.1 percent), and at least one interstate or major highway that likely generates traffic patterns that do not accurately mirror residential census populations.

Similar dramatic differences across disproportionality ratios were observed for nearly half of the twenty-seven counties examined in Pennsylvania. As demonstrated in Pennsylvania, the inconsistencies across benchmarks and the resulting differences in conclusions made regarding the findings of racial disparities can be quite extreme. These comparisons illustrate that the selection and measurement of the benchmark is critical to the statistical analyses and subsequent conclusions of racial/ethnic disparities in traffic stop studies.

The example of Jefferson County, Pennsylvania, underscores the importance of the benchmarking issue because different benchmarks produce different conclusions. Unfortunately, all benchmarks have flaws, and while different benchmarks have been examined and used in court (e.g., *U.S. v. Armstrong*, 1996), in more recent cases (e.g., *Chavez v. Illinois State Police*, 2001), the courts have used additional information to offset the weaknesses of the benchmarks (i.e., information from the National Personal Transportation Survey (U.S. Department of Transportation, Federal Highway Administration, 1995)). The issue remains that all benchmarks have flaws that need to be acknowledged when considering racial profiling and no single benchmark provides a magic bullet for assessing discrimination claims.

Which benchmark is best?

One problem with using benchmarks is when courts are given statistical evidence to determine a discriminatory effect in selective enforcement cases, the interpretation of the results will likely differ based directly on the benchmark comparison used.

Table 1
Disproportionality ratios for Black drivers using different benchmarks in twenty-seven Pennsylvania counties (Engel et al., 2005)

County name	Type of benchmark				
	All stops compared to residential census population	Stops of county residents compared to residential census population	All stops compared to weighted census traffic model	Daytime stops compared to daytime observations of roadway usage	Daytime speeding stops compared to daytime speeding observations
Allegheny	0.8	1.0	0.7	3.5	3.5
Bedford	45.6	1.9	0.7	1.1	1.0
Bucks	3.4	1.4	0.6	1.3	1.3
Centre	1.6	0.5	0.3	2.9	5.4
Chester	2.0	1.4	0.9	2.3	2.3
Clarion	13.6	0.9	0.5	1.7	1.0
Clinton	21.4	1.8	0.5	1.4	0.6
Columbia	16.9	2.4	0.6	4.4	2.6
Dauphin	0.5	0.6	0.5	3.8	2.9
Delaware	1.4	1.0	0.9	1.5	0.9
Erie	1.2	0.7	0.5	1.5	6.4
Franklin	4.2	1.8	0.6	6.7	13.2
Fulton	27.9	1.2	0.8	1.4	1.2
Indiana	1.8	0.8	0.3	3.0	3.4
Jefferson	69.2	3.7	0.3	1.2	1.0
Juniata	22.2	4.4	0.3	3.0	1.0
Lackawanna	6.4	2.9	0.5	2.0	1.0
Lehigh	3.0	2.0	0.8	2.3	1.8
McKean	0.5	0.1	0.1	1.8	2.3
Mercer	2.3	0.9	0.6	3.8	2.5
Montgomery	2.0	1.4	0.8	2.6	2.0
Montour	16.2	0.0	0.6	2.6	2.0
Susquehanna	40.8	3.1	0.5	2.2	1.2
Tioga	5.2	0.8	0.2	2.6	1.2
Washington	2.2	1.6	0.6	1.8	1.4
Westmoreland	4.8	1.9	0.7	3.8	4.7
York	3.0	1.4	0.8	1.7	1.3

As noted above, the validity and reliability of benchmarking techniques vary tremendously. While there is some consensus in the research community that residential census populations are the least reliable of the benchmarks available, there is no such consensus regarding the validity of other techniques. Lamberth (2004) argued that traffic observations are the most valid comparisons and noted that the validity and reliability of alternative forms of benchmarking are often determined by comparing the findings to data collected through traffic observations. Lamberth (2004) also noted that traffic observation and violator surveys are the only benchmarking techniques that have been identified in a court of law as an acceptable methodology for estimating driving populations, and therefore represent the most reliable and valid technique.

The key consideration for assessing the validity of benchmarks is the accuracy with which these benchmarks reflect motorists' *risks* of being stopped by police. The risk of being stopped for motorists can be affected by at least six factors: (1) where they drive, (2) when they drive, (3) how often they drive, (4) what they drive, (5) how they drive, and (6) who they are. That is, an accurate benchmark must take into consideration driving location, time of travel, driving quantity, vehicle types and conditions, driving behavior, and drivers' characteristics (Engel et al., 2004; Engel et al., 2005). All of these factors are believed to have the potential to influence motorists' likelihood of being stopped for traffic offenses, and therefore must be measured to assess similarly situated people for purposes of accurate statistical comparisons. Unfortunately, none of the available data generated by the benchmarking techniques identified above have adequately addressed all of the risk factors associated with the likelihood of motorists being stopped by police.

Violator surveys, however, have routinely been used in litigation in an effort to demonstrate discriminatory effect in selective enforcement cases (e.g., *State of New Jersey v. Ballard,* 2000; *State of New Jersey v. Clark,* 2001; *State of New Jersey v. Francis,* 2001; *State of New Jersey v. Soto et al.,* 1996; *U.S. v. Alcaraz-Arellano,* 2004; *U.S. v. Duque Nava,* 2004; *U.S. v. Mesa-Roche,* 2003). While this technique has the *potential* to measure each of the six risk factors involved in traffic stops, in practice, the data collected from most violator surveys fall far short of this potential. The methodological limitations of this data collection technique are reviewed below.

Lamberth (1994, 1996) conducted the first violator observation studies for use in selective enforcement litigation by examining speeding behavior during the mid-1990s in New Jersey and Maryland. In order to determine who was speeding on the selected roadways, trained observers rode in a vehicle traveling at the exact posted speed limit in Maryland and at five miles per hour over the posted speed limit in New Jersey. These observers recorded the race of the drivers in the cars that passed them (the speeders) as well as the drivers in cars that their vehicle passed (the non-speeders). Using this technique,

Lamberth reported that the overwhelming majority of drivers (98 percent and 93 percent in New Jersey and Maryland, respectively) were violating the posted speed limits. He also concluded that there were no significant differences in the violating behavior of Caucasian and Black drivers. The findings from these studies were used in several New Jersey state court cases to establish discriminatory effect for selective enforcement (*State of New Jersey v. Ballard,* 2000; *State of New Jersey v. Clark,* 2001; *State of New Jersey v. Francis,* 2001; *State of New Jersey v. Soto et al*., 1996).

This data collection technique, however, was fundamentally flawed. Although violator surveys, in principle, can measure all of the six factors known to influence motorists' risk of being stopped by police, Lamberth's data collection technique failed to adequately measure three factors (i.e., what motorists drive, how motorists drive, and other motorists' demographic characteristics).[6] Most important of these limitations, Lamberth's methodology did not truly assess "violators" at risk of being stopped because only a simple dichotomy of speeding or not speeding was measured. Many police agencies have formal policies or informal norms regarding the level of speeding that merits a traffic stop, warning, and/or citation, and further, these norms vary within police agencies and across jurisdictions.[7] Therefore, motorists' risk based on "how they drive" must measure more than simply "speeding or not speeding." Specifically, the *severity* and *location* of the traffic offense must be measured to assess the true risk of being stopped.

The lack of a speeding severity measure simply does not capture drivers' real risk of being stopped for that behavior even at five miles per hour over the speed limit. Furthermore, as researchers have noted, drivers differ in their levels of "speeding savvy," which suggests that some drivers may speed in ways that minimize their risks of being detected and stopped by police (e.g., motorists that travel close to other vehicles or near tractor trailers, have radar detectors, routinely slow down with traffic, etc.) (W. R. Smith et al., 2003). Therefore, motorists' risks of being stopped for speeding are not fully captured through Lamberth's methodology. Methods to determine drivers' risk of being stopped for speeding would have to rely on the same techniques for detection of speeding as the police use (Engel et al., 2004; Lange, Blackman, & Johnson, 2001). Furthermore, there are numerous additional reasons for traffic stops (e.g., other moving violations, equipment violations, etc.) that this type of data collection technique does not reliably capture. Therefore, the analyses comparing traffic stops to violator surveys must include only those traffic stops made for speeding violations.

Some scholars have suggested that there is no evidence available that shows racial groups differ in their violations of traffic laws, making them more or less eligible to be stopped by police (e.g., Lamberth, 2004; Solop, 2004a, 2004b). A growing body of research, however, does indicate that racial and ethnic differences in driving behaviors exist. Research in the travel, transportation, and accident analysis literatures shows considerable racial and ethnic differences in a variety of driving-related behaviors including seat belt usage, possession of driver's license/driving without license, fatal accident involvement, alcohol-related accident involvement, and driving under the influence (Abdel-Aty & Abdelwahab, 2000; Baker, Braver, Chen, Pantula, & Massie, 1998; Braver, 2003; Caetano & Clark, 2000; Campos-Outcalt, Prybylski, Watkins, Rothfus, & Dellapenna, 1997; Centers for Disease Control and Prevention, 2000; Chu, Polzin, Rey, & Hill, 2000; Engel et al., 2004; Everett et al., 2001; Glassbrenner, 2003; Harper, Marine, Garrett, Lezottte, & Lowenstein, 2000; Jones & Lacey, 1998; Lange et al., 2001; Lerner et al., 2001; Missouri Department of Health, 1998; Nachiondo, Robinson, & Killen, 1996; Polzin, Chu, & Rey, 2000; Royal, 2000; Schiff & Becker, 1996; W. R. Smith et al., 2003; Voas, Tippetts, & Fisher, 2000; Voas, Wells, Lestina, Williams, & Greene, 1998; Wells, Williams, & Farmer, 2002). Furthermore, four studies examining racial/ethnic differences in speeding behavior by measuring the exact speed of motorists have reported that Black drivers are significantly more likely than Caucasians to exceed the speed limit, and the actual amount over the limit at which Black motorists travel is significantly higher compared to Caucasians (Engel et al., 2004; Engel, Frank, Tillyer, & Klahm, 2006; Lange et al., 2001; W. R. Smith et al., 2003). In addition, motorists can be legitimately stopped for vehicle violations unrelated to their driving behaviors (e.g., equipment violations). Given the unequal distribution of wealth for different racial/ethnic groups in society, it is possible that minorities are more likely to drive vehicles with equipment violations due strictly to economic conditions. Together, these research findings suggest that drivers' behavior may at least partially account for racial/ethnic disparities in police stops and stop outcomes (e.g., citations and arrests).

Despite the limitations of the violator survey methodology, it was accepted in the New Jersey state courts as a reliable and valid indicator of similarly situated persons for establishing discriminatory effect in selective enforcement cases. More recently, however, other court cases have noted problems with this methodology, including: (1) observations sample sizes were too small to make meaningful generalizations, (2) observations were too limited in time (i.e., did not account for seasonal changes in driving population), (3) questionable reliability of observers' perceptions of race, particularly due to nighttime observations and because inter-rater reliability was only tested for a percentage of the observations, and (4) failure to measure the true violator population. These courts have stated that the research provided did not establish the appropriate level of social scientific evidence to establish a discriminatory effect (see *U.S. v. Alcaraz-Arellano,* 2004; *U.S. v. Duque Nava,* 2004; *U.S. v. Mesa-Roche,* 2003). Thus, data generated from the same research methodologies have been accepted as credible evidence demonstrating discrimination in some courts and not others. Further, there is wide variation in the type and amount of statistical evidence provided to courts, and as a result, great variation in court decisions based on that evidence.

In summary, while it is likely that traffic violator surveys provide a more reliable and valid estimate of roadway usage than many other techniques, this benchmark still has major limitations when providing estimates of similarly situated persons for selective enforcement litigation. Thus, to date, no benchmark has been developed that adequately addresses all of the risk factors involved in traffic stops that would identify similarly situated persons for statistical comparisons to stopped populations. That is, no data collected from benchmark methodologies currently available have fully and accurately estimated the driving

population at risk for police traffic stops. Thus, social scientists are currently unable to determine the true population of "similarly situated persons" who were eligible for traffic stops by police. They can only provide estimates, and as demonstrated previously, these estimates often vary dramatically from one another. Further, there is currently no consensus among social scientists regarding which of the benchmarking techniques (if any) are most valid.

How much statistical evidence is necessary to demonstrate discriminatory effect?

The use of statistics for the determination of discriminatory purpose and/or effect typically involves a two-step process. First, the data are analyzed using a statistical test to determine significant differences between the expected outcome (based on the benchmark), and the observed outcome (based on police traffic stop data). The most common approach in racial profiling litigation has been the use of standard deviations to assess the level of statistical evidence. As noted previously, this approach has been used in numerous jury selection cases argued in federal and state proceedings. Social scientists have recently applied this same technique to document racial and ethnic disparities for selective enforcement litigation. As will be documented in detail below, however, there are a number of problems with using standard deviation comparisons for racial profiling research.

The second step in the process involves a determination of *how much* disparity exists. Social scientists have used a number of different techniques in an attempt to determine the *amount* of racial/ethnic disparity in police stops. As will be argued below, the best method is likely the disproportionality ratio, which provides a substantive interpretation of the differences between the minority group and the Caucasian group. Unfortunately, however, this two-stage process of measuring statistical and substantive significance still cannot fully address the legal questions in selective enforcement cases. The most common techniques used by social scientists for determining both statistical and substantive significance are further described and critiqued below.

Statistical significance

A review of selective enforcement cases suggests that there is no consistency in this litigation regarding the amount of statistical evidence necessary to establish discrimination. Although the U.S. Supreme Court established the use of standard deviations to assess the level of statistical evidence in *Castaneda*, this method could generate arbitrary conclusions under one or more of the following conditions: (a) when the sampled cases are not selected independently of each other, both within and between race/ethnic groups, (b) when samples are large, (c) when sample sizes differ across studies, and (d) when the groups being compared are not internally homogeneous on race/ethnicity.

Beginning with (a) above, the U.S. Supreme Court in *Castaneda v. Partida* (1977) stated:

> The measure of the predicted fluctuations from the expected value is the standard deviation, defined for the binomial distribution as the square root of the product of the total number in the sample (here 870) times the probability of selecting a Mexican-American (0.791) times the probability of selecting a non-Mexican-American (0.209). Thus, in this case the standard deviation is approximately 12. As a general rule for such large samples, if the difference between the expected value and the observed number is greater than two or three standard deviations, then the hypothesis that the jury drawing was random would be suspect to a social scientist. (p. 497)

In algebraic notation, the formula for the standard deviation is $\sqrt{(N)(p)(q)}$, where N represents the total number of cases in a sample, p is the known proportion of all cases in the population that belong to a particular group of interest (such as Mexican-Americans), and q is the proportion of all cases in the population that do not belong to the first group (non-Mexican-Americans). Relatedly, the number of cases in a sample that *should* fall into the first group, assuming no bias, is equal to Np. With this information, a region of values can be created that range from Np to $Np+3\sqrt{(N)(p)(q)}$, with the maximum value reflecting a point estimate falling three standard deviations above Np. If the actual number of sampled cases falling into the first group is larger than this expected value, the difference is treated as "real" (i.e., statistically significant), and one concludes that there was racial/ethnic disparity in the selection process *beyond* what might be expected by chance alone.

The values of p and q represent probabilities that a sampled case *should* fall in one of the two groups, under the assumption of no disparity in the selection process. The validity of this assumption is questionable, however, because it depends on a criminal justice actor to make decisions that are completely independent from one case to the next. When selecting jurors, for example, the assumption implies that any eligible juror has the same selection probability as any other so that, in the end, the values of p and q will be reflected in the pool of selected jurors. Equal selection probabilities cannot exist, however, because the odds of selecting any one eligible juror immediately increase as soon as attorneys use up their peremptory challenges. Unless it can be assumed that the list of possible jurors is ordered randomly on race/ethnicity, which is unlikely based on a host of socio-demographic correlates of race/ethnicity (such as proportions falling into particular age groups, occupations, employment statuses, surnames, etc.), there is the potential for selection bias. Since the estimate of $\sqrt{(N)(p)(q)}$ (standard deviation) is based on equal selection probabilities, the application of this formula is questionable. The bias is further compounded by other issues such as potentially biased estimates of p and q (i.e., the pool of possible jurors might not actually include p Mexican-Americans and q non-Mexican-Americans), and the questionable assumption that an attorney's behavior towards any one eligible juror is not influenced by any decisions made up to that point in the selection process.

This same issue arises in police officer decision-making because officers conduct more than one traffic stop. When analyzing traffic stops, each officer is responsible for multiple traffic stops which introduces the possibility of correlated error *within* each officer's case load because, for example, these cases are drawn from the

same geographic area patrolled by an officer. Recent statistical techniques (i.e., multilevel analyses) have been developed to address this issue and should be considered in future traffic stop analyses. These multilevel analyses avoid pooling different units of analysis by "nesting" stops at level-1 within officers at level-2 to adjust for any correlated error across traffic stops made by particular officers (also discussed below).

Applying the statistical formula to a determination of racial/ethnic disparities in traffic stops is even more suspicious when considering all of the possible factors that interfere with the assumptions of unbiased estimates of p and q and equal selection probabilities for all drivers "eligible" to be stopped. First, the estimates of p and q are based on the racial composition of licensed drivers in a particular geographic area, and so these estimates would be potentially biased by any race/ethnic differences in travel patterns (e.g., by job and school locations, job shift, weekday versus weekend activities, etc.), incidence or prevalence of unlicensed driving, and the areas under actual surveillance at particular points in time. Second, the selection probabilities for any one eligible driver could be unequal across race/ethnic groups due to potential group differences in the incidence or prevalence of motor vehicle accidents, histories of police contacts, an officer's familiarity with particular drivers, and the population composition of areas under surveillance at particular points in time.

In summary, cases might not be independent of one another in samples of traffic stops. Officers make multiple stops within their patrol areas, which means that each officer's case load may involve stops that look more "similar" to each other in terms of driver characteristics and reasons for being stopped (e.g., some officers patrol areas more heavily populated by minorities, some are more adept at spotting expired tags, and some have more or less tolerance for driving five miles over the speed limit). When analyses are conducted at the stop level, therefore, concerns of correlated error may exist.

The issue of sample size (b) above is more applicable to extant studies of racial disparities in decision-making by *multiple* legal agents as opposed to studies of decisions made by a single actor, although sample size should be an important consideration in both types of studies. In short, the binomial distribution is more appropriately applied to small samples (Weisburd & Britt, 2004). While this actually makes sense in a study of jury selection (e.g., N = 12), the application is questionable in studies of traffic stops involving multiple officers and samples with thousands of cases. Weisburd and Britt (2004) observed that the binomial distribution is approximated by a normal curve when $Np > 5$ and $p < q$ (or when $Nq > 5$ and $q < p$, as was the situation in *Castaneda*). When compared to other probability distributions, the normal curve provides a more liberal criterion for establishing a significant difference between the observed and expected values under consideration. In other words, the number of sampled cases falling into group p is more likely to exceed the value of $Np+3\sqrt{(N)(p)(q)}$ when the normal curve is applied instead of a more platykurtic (i.e., flatter) distribution.

An obvious counter argument to the above is that the distinction between a normal distribution (Z) and a flatter distribution (t) is trivial when N is large, but the validity of such an argument relies heavily on meeting the assumption of equal selection probabilities, as described above. The correlated error that is likely to occur in studies of traffic stops could generate biased frequencies that are more likely to fall within the critical region of a normal distribution. For example, if minorities are disproportionately overrepresented among drivers traveling a stretch of road under surveillance by an exceptionally ambitious officer, the number of minorities stopped by that officer will necessarily exceed the number expected based on the estimate of p. The placement of that particular officer increases the odds that *any* driver will be stopped in his or her jurisdiction, yet minorities end up being overrepresented in the final tally based on the population composition of that officer's patrol area. Therefore, the sample of traffic stops generated from multiple police officers should not be examined as a single probability sample, but rather as a multistage cluster sample (with stops nested within officers). Each officer might be extremely uniform in their procedures, regardless of a driver's race, yet differences in stop probabilities *across* officers could more easily generate race group differences that are deemed "significant" when the normal curve is applied. In short, it is questionable that a hypothetical distribution of study outcomes involving samples compiled by different officers is even symmetrical, let alone normal.

Analyses of the decisions made by a single officer are not subject to the problem of unequal selection probabilities of cases compiled *across* officers, although it remains questionable that all of the decisions made by a single officer are truly independent of one another (for reasons noted above). Also, analyses of individual officers still might involve hundreds of cases and so the potential problem associated with applying a normal probability distribution and a more liberal criterion for establishing significant differences apply to such analyses, even if the problem is not as extreme as studies of much larger samples.

Issue (c) above relates more to a problem with hypothesis testing in general rather than any unique problem with the application of the binomial distribution. Hypothesis test formulas are purposely designed to increase the odds that any difference between observed and expected values are "real" or "significant" as sample size (N) increases, based on the idea that sampling error is reduced as more cases are included from the population. (The extrapolation to studies of entire populations is that *any* difference, no matter how small, is a *real* difference because there is no sampling error involved.) This is why related formulas measuring dispersion necessarily incorporate sample size. The implication of applying such formulas to studies of racial profiling, however, is that two different conclusions could be drawn from two studies of the *same difference* simply because of a difference in sample size. The problem might also be compounded when applying a normal probability distribution to very large samples because the distribution provides a more liberal criterion for establishing a significant difference. Again, this is not necessarily a problem in studies of jury selection with fixed N across studies. Studies of racial profiling in traffic stops, however, vary widely in sample size due to the different foci across studies (such as stops involving one officer versus multiple officers, analyses of stops

within counties versus stops across an entire state, etc). It is easier to describe this idea with a difference of proportions test instead of a test of raw differences in absolute numbers, although the same logic applies to both. In this application, the standard deviation of p (the expected proportion of cases falling into the first group) would be $\sqrt{(p)(q)/N}$ (Weisburd & Britt, 2004). With N in the denominator, the standard deviation of the distribution decreases in value with larger N even when the values of p and q remain constant across samples. If, for example, the expected value of p was .80 and the actual value was .81, this 1 percent difference between the expected and actual values of p would be considered a "significant" difference in a sample of 20,000 traffic stops, but it would not be significant in a sample of 10,000 stops. Using the formula for the standard deviation, the sample of 10,000 stops produces a standard deviation of .004: $\sqrt{(p)(q)/N} = \sqrt{(.8)(.2)/10,000} = \sqrt{.16/10,000} = \sqrt{.000016} = .004$.

Using the criterion that the actual proportion must fall at least three standard deviations above the expected proportion to be considered significantly different, one sees that the value of .81 does not meet this criterion: $p + 3\sqrt{(p)(q)/N} = .80 + (3)(.004) = .80 + .012 = .812$ (>.81).

By contrast, the sample of 20,000 stops produces a standard deviation of .0028: $\sqrt{(p)(q)/N} = \sqrt{(.8)(.2)/20,000} = \sqrt{.16/20,000} = \sqrt{.000008} = .0028$.

This standard deviation, in turn, is small enough to establish a significant difference between the expected and actual values of p: $p + 3\sqrt{(p)(q)/N} = .80 + (3)(.0028) = .80 + .0084 = .8084$ (<.81).

The sample of 10,000 cases is clearly a "large" sample, and it is questionable that doubling the sample size to 20,000 offers additional methodological rigor. The benefit of examining the larger sample lies only in a researcher's ability to take advantage of a mathematical nuance that permits him or her to treat a relatively weak difference as statistically meaningful. Yet, considering the data sets examined to date, studies could produce different conclusions based only on the jurisdictions under consideration. This scenario could produce higher odds of finding significant racial disparities in studies of state police departments versus metropolitan police departments or sheriff's offices simply because state police departments deal with a much higher volume of traffic stops during any fixed period of time.

The final concern to be discussed is the issue of comparing only two categories of race/ethnicity (i.e., issue (d) above). The particular application of the binomial in *Castaneda* could be considered appropriate based on the consideration of only two ethnic groups: Mexican-Americans and Caucasians. Analyses of traffic stops, however, often include more than one group, if only because of the sheer number of cases examined. Dichotomizing these groups into categories such as Caucasians and non-Caucasians becomes an overly crude distinction that ignores the methodological assumption of internal homogeneity on the characteristics of interest. Since the values of p and q each must vary at the expense of the other, a second category (q) that simply lumps all other race and ethnic groups together will generate larger differences between p and q that ultimately produce smaller values for the standard deviations [i.e., the product of mid-range proportions are larger than the product of more extreme proportions, such as (0.4)(0.6) = 0.24 versus (0.2)(0.8) = 0.16]. The smaller values of standard deviation, in turn, could produce expected differences that are easier to exceed with actual data because the original categories were not mutually exclusive.

The most common race and ethnic groups that are likely to appear in a large sample of traffic stops in many states include Caucasians, African Americans, Mexican or Spanish Americans, Asian Americans, and (perhaps) Native Americans. Greater racial and ethnic heterogeneity is more likely in studies conducted in the northeastern, western, and southwestern regions of the United States. Proportions of non-Caucasian drivers could appear quite large in some of these states based on the combination of several different ethnic groups. Separating these groups out could generate more comparable proportions across groups that might, in turn, lead to different conclusions regarding racial profiling. The methodological issue above might be resolved with the application of multinomial distributions or chi-square distributions to these data, although the assumption of equal selection probabilities within and between groups is still required for both.

The four issues described in this section raise important questions regarding the validity of applying the "standard" statistical formulas to a determination of racial profiling in traffic stops. At the very least, state courts need to consider the robustness of the current standard in such cases. Moreover, an important limitation of *any* statistical test is that even if a statistically significant difference is produced, the magnitude of actual disparity could be substantively small. In other words, although the tests can demonstrate if there is a difference in outcomes between what is expected and what is observed beyond what is randomly expected, there is no substantive interpretation for the result. The second part of the analysis addresses this limitation to specify the substantive interpretation.

Substantive significance

Traffic stop studies have generally used four different statistical methods to compare traffic stop data to benchmark data: (1) absolute differences, (2) relative differences, (3) disproportionality indices, and (4) disproportionality ratios. The strengths and limitations of these methods were described in detail by Fridell (2004) and summarized below.

The absolute difference simply compares the raw difference between the benchmark rate and the traffic stop rate. For example, if the benchmark rate of minority drivers is 5 percent and the traffic stop data rate is 10 percent, the absolute difference is 5 percent. There are three major limitations to this method. First, the value cannot be interpretable; that is, while there is a 5 percent difference between the benchmark rate and the traffic stop rate, it is unclear exactly what this substantively means. Second, the absolute value is relative to the values of the benchmark and traffic stop data rates; for example, if the value were changed to 90 percent for the benchmark and 95 percent for the traffic stop data, the absolute difference is still 5 percent, but if the values are transformed back into raw values, the actual number of drivers for each example are significantly different. Third, the absolute difference does not compare the rates of the minority group to the rates of the

Caucasian group. Recall this is of primary significance for judges who are ruling based on "similarly situated individuals."

The second method of interpreting differences between the benchmark and the traffic stop data is the relative difference between the two values. This is calculated by subtracting the traffic stop value from the benchmark value and then dividing by the benchmark rate. In the above example of 5 percent for the benchmark rate and 10 percent for the traffic stop rate, the relative difference would suggest a 100 percent difference in stops made by officers in comparison to the benchmark. While this overcomes the first and the second limitations of the absolute difference method, it does not compare the rate of minority drivers to the rate of Caucasian drivers stopped.

Third, using traffic stop data as the numerator and a benchmark as the denominator, a "disproportionality" or "disparity" index can be created. These indices are used to estimate the differences between the "actual" and "expected" rates of traffic stops for different racial, ethnic, gender, and age groups. The disproportionality index is calculated by dividing the stop data by the benchmark data. Disproportionality indices greater than 1.0 indicate that the rate of stops for particular groups are *greater than expected* in comparison to the benchmark. A disproportionality index of less than 1.0 indicates that the rates of traffic stops for particular groups are *less than expected* based on the benchmark. The larger the size of the disproportionality index, the larger the disparity between the actual and expected rate of stops. The disproportionality index still does not provide a comparison between the rates of minority drivers stopped with the rate of stops for Caucasian drivers.

The final method for calculating differences between the benchmark data and the traffic stop data is the disproportionality ratio. The interpretation of this value is the same as that of the disproportionality index (i.e., a value of 1.0 indicates no disparity; values greater than 1.0 indicate a higher likelihood of being stopped if of minority status; and values less than 1.0 indicate minorities are less likely to be stopped in comparison to Caucasian drivers). This is the only method that integrates both the rates of minorities and Caucasian drivers by comparing the differences to provide an interpretable value. In other words, both the disproportionality index for minorities and Caucasians is used in the calculation of the disproportionality ratio and the result is interpreted as how many times more likely is the minority group to be stopped in comparison to the Caucasian group. One other major advantage of the disproportionality ratio is its cross-jurisdiction comparison ability. In other words, disproportionality ratios can be directly compared across jurisdictions because they are standardized, which could be a significant strength for consistency across legal decisions.

Collectively, these statistical techniques attempt to quantify the level of racial/ethnic disparities in traffic stops. That is, their purpose is to demonstrate "how much" racial/ethnic disparity exists; however, these statistics have limitations. First, as indicated previously, these measures are based on benchmark data that likely do not capture drivers' true risk of being stopped by police. Second, only one of these measures—the disproportionality ratio—is readily interpretable and makes comparisons across racial groups. Finally, although disproportionality ratios are readily interpretable, there is no scientifically accepted level at which a measure of disparity is considered illegitimate or unjustified. For example, as described above, a disproportionality ratio value of 1.0 indicates that there is no higher likelihood for minority drivers to be stopped when compared with Caucasian drivers. As the values rise above 1.0, more disparity exists; however, it is not clear how much disparity is too much. Does a disproportionality ratio of 1.5, meaning minority drivers are 1.5 times more likely to be stopped, demonstrate too much disparity by a legal standard? What about 2.0 or 3.0, etc.? While the obvious goal would be equality between the races, the question remains whether or not a threshold value exists that indicates an unacceptable level of disparity. Given the limitations of the statistical analyses, a "bright line" value for measures of disparity should become a legal or policy decision, not a statistical one because the use of the standard deviation is fraught with difficulties. Courts employing social science statistics must be cautious that the evidence is both statistically accurate, but also substantively meaningful. Without both of these components, social science statistics can be misinterpreted and misused.

Discussion

Due to increasing pressure from politicians, the public, and policymakers, racial profiling is now a well-researched social phenomenon. The increased concern regarding profiling has led to the intersection between the legal system and social science research. As noted above, racial profiling cases are often litigated as selective enforcement cases and statistical evidence is used to establish discriminatory purpose and/or effect. This requires that social science research be able to adequately measure and compare rates of racial/ethnic groups' traffic stops to their representation in the population eligible for traffic stops. In legal terms, it requires a comparison of "similarly situated individuals." In addition, findings of discriminatory purpose and/or effect are based in part on the strength or amount of the statistical evidence demonstrated. Unfortunately, social science research has yet been unable to adequately address these two legal concerns.

From a social science perspective, one of the most important issues involved in traffic stop data analyses is the interpretation of the analyses and conclusions offered by the analyst. Claims of racial/ethnic discrimination based on traffic stop studies are problematic for a number of reasons. First, as documented above, there is no benchmark available that reliably measures all of the drivers' risks of being stopped for a traffic offense. Serious discrepancies exist in disproportionality ratios calculated using different benchmarks. While a multitude of methods are available to researchers, no single benchmarking technique is free from limitations; thus, there is no agreed upon measure of "similarly situated individuals" within the driving population that is considered the most appropriate by social scientists.

Second, despite the holding in *Castaneda*, there is no acceptable level of disparity that has been recognized by social science researchers that can be said to accurately demonstrate police discrimination. The use of standard deviation comparisons as accepted by the court in *Castaneda* does not appear to be an appropriate technique for analyzing racial profiling data.

Consequently, other methodological and/or statistical techniques need to be applied to improve the social scientists' ability to report accurately to the court.

Third, and perhaps most importantly, even if a reliable benchmark was created and compared to traffic stop data, and social scientists agreed on an appropriate threshold value of disparity, the demonstration of racial disparities in traffic stops cannot be directly attributed to profiling because the statistical analyses do not currently measure alternative explanations of racial disparities. That is, the statistical analyses used to examine traffic stop data simply cannot determine the individual motivation of officers when making stopping decisions. Data and methodological limitations, along with ethical practicalities, prevent social scientists from pinpointing the causal factors leading to racial/ethnic disparities in traffic stops.

Racial discrimination can only be assessed if other possible legitimate reasons for observed racial disparities have been considered and ruled out. It cannot be determined with data from current traffic stop studies if the observed racial/ethnic disparities are due to discrimination, because of the inability to measure alternative factors that might account for the disparities. As noted by Engel et al. (2002, p. 250), "the problem with interpreting these findings is that the mere presence of disparity in the aggregate rate of stops does not in itself demonstrate racial prejudice, any more than racial disparity in prison populations demonstrates racial prejudice by sentencing judges." Likewise, Fridell (2004, p. 2) noted that "because the data will never 'prove' or 'disprove' racially biased policing, we contend that vehicle stop data collection and analysis should never be viewed—either by police or resident stakeholders—as a 'pass-fail test.'" Thus, measurements of the factors that influence individual officer decision-making are critical to determine whether or not officers are acting on racial prejudice, animus, cognitive bias, or profiling.

It is important to recognize that racial and ethnic disparities in traffic stops can be the result of either bias or non-bias mechanisms (Tomaskovic-Devey, Mason, & Zingraff, 2004). While definitions of "racial profiling" vary, all suggest that profiling involves some form of racial bias, whether through racial animus, prejudice, or some form of organizational discrimination. Yet, the source of racial disparities in police stops could also be the result of non-bias mechanisms (Tomaskovic-Devey et al., 2004). For example, numerous scholars have noted that measuring alternative, race-neutral factors, including racial differences in driving patterns, location, frequency, and/or degree of law-violating behavior, as well as spatial characteristics such as high police presence, might be considered legitimate (i.e., non-bias) explanations of racial/ethnic disparities (e.g., Alpert Group, 2004; Cordner, Williams, & Zuniga, 2001; Cox, Pease, Miller, & Tyson, 2001; Criminal Justice Training Commission, 2001; Engel et al., 2004; Farrell et al., 2004; Farrell et al., 2003; Fridell, 2004; Lansdowne, 2000; Novak, 2004; Rojek et al., 2004; M. R. Smith & Petrocelli, 2001; W. R. Smith et al., 2003; Texas Department of Public Safety, 2001; Tomaskovic-Devey et al., 2004; Withrow, 2004; Zingraff et al., 2000).

The aforementioned difficulties associated with measuring "similarly situated" individuals leads to a significant concern for using social science research within the legal domain. As a result, social science research and its use in legal decisions would be more appropriately applied to consideration of post-stop outcomes. A focus on this component of the police-citizen interaction is statistically and methodological on firmer ground in regard to making a determination as to the existence of racial bias. This is not to suggest that social science can unequivocally determine if post-stop outcomes are as a result of racial bias, but social science can provide an improved understanding of the factors associated with the likelihood of a citizen being warned, cited, arrested, or searched.

Research on post-stop outcomes has a distinct advantage over studying potential biases in traffic stops because the benchmarking limitations that hinder stop analyses are not present for post-stop analyses. As previously described, to determine if there is racial disparity in traffic stops, development of a benchmark is necessary to allow a comparison between who has been stopped and those "expected" to be stopped if no bias existed. This requirement is not necessary for post-stop outcomes, as the data collected for analyses contains all possible outcomes (i.e., a warning, citation, arrest, search, or no formal response). As a result, the limitations of benchmarking do not apply to post-stop outcomes and are thus more amenable to analyses by social science. Furthermore, due to the fact that information is available on the entire scope of post-stop outcomes, statistical significance testing and multivariate analyses are able to be used to assist in identifying patterns of disparity in post-stop outcomes. These types of analyses are not amenable to traffic stops due to not having information on every driver who has the potential to be stopped (i.e., a necessary component for statistical analyses).

Notwithstanding social sciences' improved ability to make determinations of racial bias in post-stop outcomes, measuring all factors associated with an officer's decision to warn, cite, arrest, or search is limited. To improve the understanding of unmeasured influences in post-stop outcomes, future research on racial profiling needs to focus on these factors. Two separate, yet related, approaches could be employed by social scientists to overcome the aforementioned weaknesses in data collection and analyses. First, improvements in data collection would assist in addressing the breadth of knowledge regarding post-stop outcomes. To address this goal, data collection efforts must be valid and thorough; that is, there needs to be internal mechanisms in place to ensure that all information on post-stop outcomes is gathered and accurate. With multiple jurisdictions currently mandated to collect data on all police-citizen encounters, agencies and research teams involved in data collection must ensure that the validity of the data is maintained through internal auditing mechanisms (for summary, see Fridell, 2004). It is imperative that all information on post-stop outcomes is collected to ensure that missing information does not bias the results of the analyses.

Furthermore, data collection efforts should include a comprehensive list of variables to ensure that the most relevant explanatory factors associated with officer decision making are gathered. Improving data collection efforts by examining all measurable factors will assist in developing a holistic understanding of post-stop outcomes. Such additional, not routinely collected, variables may include, but are not limited to the

characteristics of the driver and/or the vehicle. For example, the driver's demeanor and/or previous record may be associated with the likelihood of being searched. Alternatively, vehicle characteristics such as the condition of the vehicle and modifications to the vehicle may have an impact on an officer's decision to engage in a particular type of stop outcome. These hypothesized empirical relationships could be examined by the emphasis on collecting such information during the data collection process.

Another method of improving data collection would be to use interviews and focus groups with officers. Interviewing officers or conducting focus groups with officers would improve an understanding of their thought processes both prior to and during an encounter with a citizen. Information of this type could lend a great deal of insight into explaining potential biased outcomes for minority populations by identifying unmeasured influences that are associated with both pre- and post-stop officer decision making. If disparities are discovered, interviews and focus groups could highlight the source of such inconsistencies. For example, it may be that officers are not familiar with specific cultural or religious responses by minority citizens to authority (Engel & Johnson, 2006). Officers misunderstanding of citizens' behavior could be targeted in training to alleviate such discrepancies.

The second area that social science can improve is in its understanding of methodological and statistical techniques of officer decision making to ensure that the results used in court are accurate and interpretable. For example, linear and nonlinear hierarchical modeling, the outcome test, and propensity scores are all recent additions to the analytical toolbox of social scientists and are applicable to examining post-stop outcomes. Specifically, hierarchical linear and nonlinear modeling would allow researchers to partition the effects of neighborhood, county, or police jurisdictions from those individual factors related to the officer or citizen. The use of this technique has become commonplace in other areas of criminal justice research and could become an asset in identifying the specific factors associated with officer decision-making. As previously discussed, the nature of police work produces multiple police-citizen encounters by one officer, which leads to statistical problems. The use of multilevel techniques could address such shortcomings and improve the ability of social scientists to parcel out the multitude of factors associated with officer decision making. Statistical analyses that are based on sound mathematical principles and are matched to the complex nature of police-citizen encounters can be an effective tool for understanding and identifying racial bias.

In addition to the use of linear and nonlinear hierarchical modeling techniques, economists have entered the racial profiling discussion by applying the outcome test to data collected on post-stop outcomes (Knowles, Persico, & Todd, 2001). This technique has most frequently been used to examine search and seizure patterns of police departments by comparing stop outcomes across racial/ethnic groups to disentangle statistical relationships from racial bias. While the use of the outcome test in other social science arenas is accepted, some authors question its applicability to understanding officer decision-making (Engel, in press). The applicability of the outcome test to assist in identifying racial bias in officer decision making is still an open question to be determined by social scientists and the legal community.

Another statistical technique developed to analyze post-stop outcomes is the use of propensity scores. Ridgeway (2006) has argued for the use of propensity scores as an alternative to current regression techniques, which are used to parcel out the specific influences from various factors potentially associated with post-stop outcomes. In essence, propensity scores adjust for the confounding variables that are associated with outcomes such as warnings, citations, or arrests by weighting the police-citizen encounters to allow a direct comparison between Caucasian and minority drivers. Using data from Oakland, California, Ridgeway (2006) demonstrated that the racial bias existing in outcomes using traditional regression techniques is removed with the use of propensity scores. The validity of this technique as applied to traffic stop data has not yet been discussed in the social science literature. Nevertheless, developing new statistical tools is critical to assist in understanding police-citizen encounters.

The use of multilevel analyses, the outcome test, and propensity scores have only recently surfaced within the academic literature regarding racial profiling. The acceptance of such techniques has yet to be determined within the academic realm, and has not been applied consistently within the legal arena to date. These techniques have the potential to address some of the limitations described above; however, it is too early to assess if they will provide the necessary information to assist the courts in assessing claims of racial profiling.

Conclusion

In summary, for traffic stops, it is not currently possible to accurately measure "similarly situated persons" because comparisons of standard deviations are inappropriate and there is no consensus among social scientists regarding the "amount" (or the statistical technique used to determine the amount) of evidence necessary to establish discriminatory effect. Even if these issues were resolved with stronger methodological research designs and statistical analyses, researchers would still be unable to reliably determine if the racial/ethnic disparities observed are the result of officers' biases because factors that influence officers' decision making are not measured in traffic stop data that is routinely collected by police agencies. Yet, the courts and policymakers still need to determine whether or not the racial/ethnic disparities reported in traffic stops studies are based on discriminatory practices. Focusing on post-stop outcomes has significant potential to assist legal decisions if data collection and analysis techniques continue to improve. This discussion has outlined a few suggestions for improving this situation, while recognizing that a definitive answer of racial profiling is not possible with social science research. Social science research can play an important role in understanding police decision making, but its strengths need to be balanced with extreme caution against overstepping its evidentiary power in courtroom proceedings. If the courts and social scientists continue to ignore these limitations, both social science research and decisions within the legal system will be tarnished by inaccurate conclusions regarding racial/ethnic discrimination at the hands of the police.

Acknowledgements

A previous version of this article was presented at the American Society of Criminology annual meetings, Nashville, Tennessee, November 2004. The authors would like to thank Jennifer Cherkauskas and Geoffrey Alpert for their thoughtful comments.

Notes

1. For the purposes of this discussion, *disparities* refer to differences in rates or outcomes, whereas *discrimination* reflects **intentional** differential treatment in those rates or outcomes. These definitions are analogous to the distinction between discriminatory effect and discriminatory purpose (*Whren et al. v. United States*, 1996).

2. They also demonstrated that an overwhelming majority of the 171 individuals excluded based on these criteria were Black (96 percent), and effectively argued that their exclusion was likely due to racial considerations rather than morality and intelligence (*Turner et al. v. Fouche et al.,* 1970).

3. For a review of the methodological strengths and weaknesses of these benchmarks, see Engel and Calnon, 2004a; Fridell, 2004; Fridell, Lunney, Diamond, and Kubu, 2001; Walker, 2001.

4. The weighted traffic flow model was developed using residential census data and the traffic stop data collected by the officers. For the speeding observations and traffic stops, "speeding" was operationalized as ten miles per hour or more over the posted speed limit. See Engel et al., 2005 for a more thorough description of the methodologies.

5. The disproportionality ratios were calculated by dividing the minority disproportionality index by the majority disproportionality index. The resulting value is the disproportionality ratio and is interpreted as the likelihood of being stopped if you are part of the racial group of interest. Disproportionality indices for only twenty-seven of the sixty-seven counties in Pennsylvania were calculated because only these counties had all five benchmark comparisons available (i.e., roadway and speeding observations were only conducted in the twenty-seven counties identified). A disproportionality ratio value of 1.0 indicates that there is no greater likelihood of Black drivers being stopped compared to White drivers. Any value above 1.0 indicates a Black driver is more likely to be stopped compared to Whites, whereas a value below 1.0 suggests that White drivers are more likely to be stopped compared to Black drivers. For details regarding the roadway observation sampling design, see Engel and Calnon, 2004a; Engel et al., 2004; Engel et al., 2005.

6. Additional critiques of Lamberth's methodology have concluded that due to inadequate sampling designs, the other three risk factors (i.e., where, when, and how often motorists drive) have been inadequately measured as well (Wilson, 2000) (also see *U.S. v. Alcaraz-Arellano*, 2004; *U.S. v. Duque Nava*).

7. For example, the average speed at which Pennsylvania State Troopers stopped motorists in 2002-2003 was 19.1 miles per hour above the posted speed limit (Engel et al., 2004). This average, however, varied considerably across the eighty-nine different Pennsylvania State Police stations throughout the Commonwealth (e.g., compare Tionesta Station's 14.1 miles per hour average amount over the speed limit for traffic citations to Media Station's 24.8 miles per hour average amount over the speed limit for traffic citations).

References

Abdel-Aty, M. A., & Abdelwahab, H. T. (2000). Exploring the relationship between alcohol and the driver characteristics in motor vehicle accidents. *Accident Analysis and Prevention*, *32*, 473−482.

Alpert, G. P., & Dunham, R. G. (2004). *Understanding police use of force: Officers, suspects and reciprocity.* Cambridge, UK: Cambridge University Press.

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

Alpert Group. (2004). *Miami-Dade Police Department racial profiling study* (Rep. to the Miami-Dade Police Department). Miami, FL: Author.

American Civil Liberties Union. (1999). *Arrest the racism: Racial profiling in America.* Retrieved March 14, 2005, from http://www.aclu.org/profiling/index.html

Baker, S. P., Braver, E. R., Chen, L. H., Pantula, J. F., & Massie, D. (1998). Motor vehicle occupant deaths among Hispanic and Black children and teenagers. *Archives of Pediatric and Adolescent Medicine*, *152*, 1209−1212.

Bittner, E. (1970). *The functions of police in modern society.* Washington, DC: U.S. Government Printing Office.

Boyle, J., Dienstfrey, S., & Sothoron, A. (1998). *National survey of speeding and other unsafe driving actions. Vol. 2: Driver attitudes and behavior.* Washington, DC: National Highway Traffic Safety Administration.

Braver, E. R. (2003). Race, Hispanic origin, and socioeconomic status in relation to motor vehicle occupant death rates and risk factors among adults. *Accident Analysis and Prevention*, *35*, 295−309.

Caetano, R., & Clark, C. L. (2000). Hispanics, Blacks and Whites driving under the influence of alcohol: Results from the 1995 national alcohol survey. *Accident Analysis and Prevention*, *32*, 57−64.

Campos-Outcalt, D., Prybylski, D., Watkins, A. J., Rothfus, G., & Dellapenna, A. (1997). Motor-vehicle crash fatalities among American Indians and non-Indians in Arizona, 1979 through 1988. *American Journal of Public Health*, *87*, 282−285.

Centers for Disease Control and Prevention. (2000). *National vital statistics reports* (Vol. 47). Washington, DC: National Center for Disease Control.

Chu, X., Polzin, S. E., Rey, J. R., & Hill, E. T. (2000). Mode choice by people of color for non-work travel. In *Travel patterns of people of color* (Publication No. FHWA-PL-00-24, chap. 6) Washington, DC: U.S. Department of Transportation, Federal Highway Administration.

Cordner, G., Williams, B., & Zuniga, M. (2001). *Vehicle stop study: Year end report.* San Diego, CA: San Diego Police Department.

Cox, S. M., Pease, S. E., Miller, D. S., & Tyson, C. B. (2001). *State of Connecticut 2000–2001 report of traffic stops statistics.* Rocky Hill, CT: Division of Criminal Justice.

Criminal Justice Training Commission. (2001). *Report to the legislature on routine traffic stop data.* Seattle: Washington State Patrol and Criminal Justice Training Commission.

Eck, J. E., Liu, L., & Bostaph, L. (2003). *Police vehicle stops in Cincinnati.* Retrieved March 10, 2005, from http://www.cincinnatioh.gov/police/downloads_pdf6937.pdf

Engel, R.S. (in press). A critique of the "outcome test" in racial profiling research. *Justice Quarterly.*

Engel, R. S., & Calnon, J. M. (2004a). Comparing benchmark methodologies for police-citizen contacts: Traffic stop data collection for the Pennsylvania State Police. *Police Quarterly*, *7*, 97−125.

Engel, R. S., & Calnon, J. M. (2004b). Examining the influence of drivers' characteristics during traffic stops with police: Results from a national survey. *Justice Quarterly*, *21*, 49−90.

Engel, R. S., Calnon, J. M., & Bernard, T. J. (2002). Theory and racial profiling: Shortcomings and future directions in research. *Justice Quarterly*, *19*, 201−225.

Engel, R. S., Calnon, J. M., Liu, L., & Johnson, R. (2004). *Project on police-citizen contacts: Year 1, final report.* Retrieved March 5, 2005, from http://www.psp.state.pa.us/psp/lib/psp/pdf/psp_police_citizens_contact_final_report_2002−2003.pdf

Engel, R. S., Calnon, J. M., Tillyer, R., Johnson, R., Liu, L., & Wang, X. (2005). *Project on police-citizen contacts: Year 2, final report.* Retrieved June 13, 2006, from http://www.psp.state.pa.us/PSP/Lib/psp/PSP_Year_2_-Citizen_ Contact_Report.pdf

Engel, R. S., Frank, J., Tillyer, R., & Klahm, C. (2006). *Cleveland division of police traffic stop study: Final report.* Retrieved July 15, 2006, from http://www.uc.edu/criminaljustice/ProjectReports/Cleveland_Traffic_Stop_Study.pdf

Engel, R. S., & Johnson, R. (2006). Toward a better understanding of racial and ethnic disparities in search and seizure rates. *Journal of Criminal Justice*, *34*, 605−617.

Engel, R. S., & Silver, E. (2001). Policing mentally disordered subjects: A reexamination of the criminalization hypothesis. *Criminology*, *39*, 225−252.

Everett, S. A., Shults, R. A., Barrios, L. C., Sacks, J. J., Lowry, R., & Oeltmann, J. (2001). Trends and subgroup differences in transportation-

related risk and safety behaviors among high school students, 1991–1997. *Journal of Adolescent Health*, *28*, 228−234.

Farrell, A., McDevitt, J., Bailey, L., Andresen, C., & Pierce, E. (2004). *Massachusetts racial and gender profiling study.* Retrieved March 12, 2005, from http://www.racialprofilinganalysis.neu.edu/IRJsite_docs/finalreport.pdf

Farrell, A., McDevitt, J., Cronin, S., & Pierce, E. (2003). *Rhode Island traffic stop statistics act: Final report.* Boston: Northeastern University, Institute on Race and Justice.

Fridell, L. (2004). *By the numbers: A guide for analyzing race data from vehicle stops.* Washington, DC: Police Executive Research Forum.

Fridell, L., Lunney, R., Diamond, D., & Kubu, B. (2001). *Racially biased policing: A principled response.* Washington, DC: Police Executive Research Forum.

Fyfe, J. J. (1982). Blind justice: Police shootings in Memphis. *Journal of Criminal Law and Criminology*, *73*, 707−722.

Fyfe, J. J. (1988). Police use of deadly force: Research and reform. *Justice Quarterly*, *5*, 165−205.

General Accounting Office. (2000, March). *Better targeting of airline passengers for personal searches could produce better results* (Publication No. GAO/GGD-00-38). Washington, DC: General Accounting Office, U.S. Customs Service.

Glassbrenner, D. (2003). *Safety belt use in 2002 – Demographic characteristics.* Washington, DC: National Highway Traffic Safety Administration.

Harper, J. S., Marine, W. M., Garrett, C. J., Lezottte, D., & Lowenstein, S. R. (2000). Motor vehicle crash fatalities: A comparison of Hispanic and non-Hispanic motorists in Colorado. *Annals of Emergency Medicine*, *36*, 589−596.

Harris, D. A. (1999). *Driving while Black: Racial profiling on our nation's highways.* Retrieved March 11, 2005, from http://www.aclu.org/profiling/report/index.html

Harris, D. A. (2002). *Profiles in injustice: Why racial profiling cannot work.* New York: The New Press.

Jones, R. K., & Lacey, J. H. (1998). *Alcohol highway safety: Problem update.* Washington, DC: National Highway Traffic Safety Administration.

Klinger, D. (1994). Demeanor or crime? Why "hostile" citizens are more likely to be arrested. *Criminology*, *32*, 475−493.

Knowles, J., Persico, N., & Todd, P. (2001). Racial bias in motor vehicle searches: Theory and evidence. *Journal of Political Economy*, *109*, 203−299.

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 on interchanges 1 and 3 from the years 1988 through 1991.* Unpublished manuscript, Temple University, Philadelphia.

Lamberth, J. (1996). *A report to the ACLU.* New York: American Civil Liberties Union.

Lamberth, J. (2003). *Racial profiling study and services: A multijurisdictional assessment of traffic enforcement and data collection in Kansas.* Retrieved March 20, 2005, from http://www.racialprofilinganalysis.neu.edu/IRJ_docs/KS_2003.pdf

Lamberth, J. (2004, July). *Observation benchmarking.* Paper presented at the meeting of By the Numbers: How to Analyze Race Data from Vehicle Stops (sponsored by the Police Executive Research Forum and COPS), Las Vegas, NV.

Langan, P. A., Greenfeld, L. A., Smith, S. K., Durose, M. R., & Levin, D. J. (2001). *Contacts between police and the public: Findings from the 1999 national survey* (NCJ 184957). Washington, DC: U.S. Department of Justice, Bureau of Justice Statistics.

Lange, J. E., Blackman, K. O., & Johnson, M. B. (2001). *Speed violation survey of the New Jersey Turnpike: Final report.* Trenton, NJ: Office of the Attorney General.

Lange, J.E., & Voas, R.B. (2000). *Survey of drivers on the New Jersey Turnpike: Final report*. Trenton, NJ/Washington, DC: Office of the Attorney General/U.S. Department of Justice, Civil Rights Division, Special Litigation.

Lansdowne, W. M. (2000). *Vehicle stop demographic study.* San Jose, CA: San Jose Police Department.

Lempert, R. O. (1989). Humility is a virtue: On the publicization of policy-relevant research. *Law and Society Review*, *23*, 145−161.

Lerner, E. B., Jehle, D. V., Billittier, A. J., Moscati, R. M., Connery, C. M., & Stiller, G. (2001). The influence of demographic factors on seatbelt use by adults injured in motor vehicle crashes. *Accident Analysis and Prevention*, *33*, 659−662.

Mastrofski, S. D., Worden, R. E., & Snipes, J. B. (1995). Law enforcement in a time of community policing. *Criminology*, *33*, 539−563.

Missouri Department of Health. (1998). *Mortality by race and gender.* Jefferson City: Missouri Department of Health and Social Services.

Nachiondo, J. M., Robinson, T. N., & Killen, J. D. (1996). Do ethnicity and level of acculturation predict seatbelt use in adolescents? *Pediatrics Research*, *39*, 7.

National Research Council. (2004). *Fairness and effectiveness in policing: The evidence.* Washington, DC: National Academies Press.

Novak, K. J. (2004). Disparity and racial profiling in traffic enforcement. *Police Quarterly*, *7*, 65−96.

Novak, K. J., Frank, J., Smith, B. W., & Engel, R. S. (2002). Revisiting the decision to arrest: Comparing beat and community officers. *Crime and Delinquency*, *48*, 70−98.

Pickrell, D., & Schimek, P. (1998). *Trends in personal motor vehicle ownership and use: Evidence from the nationwide personal transportation survey.* Unpublished manuscript, U.S. Department of Transportation Volpe Center, Cambridge, MA.

Polzin, S. E., Chu, X., & Rey, J. R. (2000). Demographics of people of color: Findings from the nationwide personal transportation survey. In *Travel patterns of people of color* (Publication No. FHWA-PL-00-24, chap. 2) Washington, DC: U.S. Department of Transportation, Federal Highway Administration.

Ramirez, D., McDevitt, J., & Farrell, A. (2000). *A resource guide on racial profiling data collection systems: Promising practices and lessons learned.* Washington, DC: U.S. Department of Justice.

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

Riksheim, E., & Chermak, S. M. (1993). Causes of police behavior revisited. *Journal of Criminal Justice*, *21*, 353−382.

Rojek, J., Rosenfeld, R., & Decker, S. (2004). The influence of driver's race on traffic stops in Missouri. *Police Quarterly*, *7*, 126−147.

Royal, D. (2000). *Racial and ethnic group comparisons: National surveys of drinking and driving attitudes and behavior – 1993, 1995, and 1997.* Washington, DC: National Highway Traffic Safety Administration.

Rubinstein, J. (1973). *City police.* New York: Ballantine.

Schiff, M., & Becker, T. (1996). Trends in motor vehicle traffic fatalities among Hispanic, non-Hispanic, Whites, and American Indians in New Mexico, 1958–1990. *Ethnic Health*, *1*, 283−291.

Sherman, L. W., & Berk, R. A. (1984). The specific effects of arrest for domestic assault. *American Sociological Review*, *49*, 261−272.

Sherman, L. W., & Cohen, E. G. (1989). The impact of research on legal policy: The Minneapolis domestic violence experiment. *Law and Society Review*, *23*, 117−144.

Skolnick, J. H. (1966). *Justice without trial: Law enforcement in democratic society.* New York: Wiley.

Smith, M. R., & Alpert, G. (2002). Searching for direction: Courts, social science, and the adjudication of racial profiling claims. *Justice Quarterly*, *19*, 673−703.

Smith, M. R., & Petrocelli, M. (2001). Racial profiling? A multivariate analysis of police traffic stop data. *Police Quarterly*, *4*, 4−27.

Smith, W. R., Tomaskovic-Devey, D., Zingraff, M. T., Mason, H. M., Warren, P. Y., & Wright, C. P. (2003). *The North Carolina highway traffic study.* Retrieved March 3, 2005, from http://www.ncjrs.gov/pdffiles1/nij/grants/204021.pdf

Solop, F. I. (2004a). *Comparative statistical analysis of I-17 stop data and I-17 violator data: Yavapai County, Arizona*. Unpublished manuscript.

Solop, F. I. (2004b). *Statistical analysis of I-40 stop data and I-40 violator data: Coconino County, Arizona*. Unpublished manuscript.

Terrill, W., & Mastrofski, S. D. (2002). Situational and officer-based determinants of police coercion. *Justice Quarterly*, *19*, 215−248.

Texas Department of Public Safety. (2001). *Traffic stop data report.* Retrieved March 14, 2005, from http://www.txdps.state.tx.us/director_staff/public_information/trafrep2001totals.pdf

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

Tonry, M. (1995). *Malign neglect.* New York: Oxford University Press.
U.S. Department of Transportation, Federal Highway Administration (1995). *Nationwide personal transportation survey* (ICPSR version) [Data file]. Ann Arbor, MI: Inter-university Consortium for Political and Social Research.
Van Maanen, J. (1974). A developmental view of police behavior. In H. Jacob (Ed.), *The potential for reform of criminal justice.* Beverly Hills, CA: Sage.
Voas, R. B., Tippetts, A. S., & Fisher, D. A. (2000). *Ethnicity and alcohol-related fatalities: 1990 to 1994.* Washington, DC: National Safety Administration.
Voas, R. B., Wells, J., Lestina, D., Williams, A., & Greene, M. (1998). Drinking and driving in the United States: The 1996 national roadside survey. *Accident Analysis and Prevention*, *30*, 267−275.
Walker, S. E. (1993). *Taming the system: The control of discretion in criminal justice, 1950–1990.* New York: Oxford University Press.
Walker, S. E. (2001). Searching for the denominator: Problems with police traffic stop data and an early warning system solution. *Justice Research and Policy*, *3*, 63−95.
Weisburd, D., & Britt, C. (2004). *Statistics in criminal justice* (2nd ed.). Belmont, CA.: Thomson and Wadsworth.
Wells, J. K., Williams, A. F., & Farmer, C. F. (2002). Seat belt use among African Americans, Hispanics and Whites. *Accident Analysis and Prevention*, *34*, 523−529.
Wilson, J. (2000). *Analysis of motorists stops on I-40 by race/ethnic origin* (Rep. prepared for the U.S. Attorney General Office, District of Arizona). Unpublished manuscript.
Withrow, B. L. (2004). Driving while different: A potential theoretical explanation for race-based policing. *Criminal Justice Police Review*, *15*, 344−364.
Worden, R. (1989). Situational and attitudinal explanations of police behavior: A theoretical reappraisal and empirical assessment. *Criminology*, *31*, 203−241.
Zingraff, M. T., Mason, H., Smith, W., Tomaskovic-Devey, D., Warren, P., McMurray, H., et al. (2000). *Evaluating North Carolina State Highway Patrol data: Citations, warnings, and searches in 1998.* Raleigh: North Carolina Department of Crime Control and Public Safety.

Cases cited

Alexander v. Louisiana, 405 U.S. 625 (1972).
Castaneda v. Partida, 430 U.S. 482 (1977).
Chavez v. Illinois State Police, 310 F.3d (7th Cir. 2001).
McCleskey v. Kemp, 481 U.S. 279 (1987).
State of New Jersey v. Ballard, 752 A. 2d 735 (2000).
State of New Jersey v. Clark, 785 A. 2d 59 (2001).
State of New Jersey v. Francis, 775 A. 2d 79 (2001).
State of New Jersey v. Soto et al., 734 A. 2d 350 (N.J. Super. 1996).
Turner et al. v. Fouche et al., 396 U.S. 346 (1970).
U.S. v. Alcaraz-Arellano, 302 F. Supp. 2d 1217 (2004).
U.S. v. Armstrong, 517 U.S. 456 687 (1996).
U.S. v. Duque Nava, 315 F. Supp. 2d 1144 (2004).
U.S. v. Mesa-Roche, 288 F. Supp. 2d 1172 (2003).
Whren et al. v. United States, 517 U.S. 806 (1996).
Wilkins v. Maryland State Police et al., Civ. No. MJG-93-468 (D.Md. 1993).
Wo v. Hopkins, 118 U.S. 356 (1886).