Comparing Benchmark Methodologies for Police-Citizen Contacts:
Traffic Stop Data Collection for the Pennsylvania State Police

ROBIN SHEPARD ENGEL
University of Cincinnati

JENNIFER M. CALNON
The Pennsylvania State University

The political and social pressure for police departments to collect race-based traffic and pedestrian contact information has led to the accumulation of abundant sources of police-citizen contact data. Many of the current data collection efforts, however, do not include accurate benchmarks for data comparisons. The strengths and limitations of the six most prominent benchmark measures used in current studies of police-citizen contacts—census data, observations of roadway usage, official accident data, assessments of traffic violating behavior, citizen surveys, and internal departmental comparisons—are described. Focusing on the Project on Police-Citizen Contacts, a large-scale data collection effort of traffic stops for the Pennsylvania State Police, four different benchmarks for statewide comparisons are described. The continued improvement of data collection efforts at the local and state levels through the use of multiple benchmark measures is discussed.

Keywords: racial profiling; benchmarks; traffic stops; police-citizen contacts; Pennsylvania State Police

The decision to collect race-based information for police contacts with citizens is a delicate one for police administrators. The methodological and empirical issues involved in data collection and interpretation are numerous.
and complex (for a review, see Ramirez, McDevitt, & Farrell, 2000). Researchers have struggled with issues surrounding proper benchmarks and whether detected differences reflect racial discrimination by police officers (Engel, Calnon, & Bernard, 2002). Comparisons enabled by the data are often inappropriate and easily misinterpreted by researchers, police administrators, and the public. Nevertheless, the political and social pressure for police departments to collect citizen race-based information has led to dramatic increases in both effort and expenditures nationwide by police departments and has resulted in the accumulation of abundant sources of police-citizen contact data.

Yet the fact remains that the data collected by police officials must be compared against some reasonable criteria. The problem is that much of the traffic and pedestrian stop data collected does not include accurate comparisons (Engel et al., 2002). Although many police departments are collecting this information, police administrators are now faced with concerns of how to accurately interpret and use the data. As described by one official from the Philadelphia Police Department during testimony at a state legislative hearing: “We have a lot of data, but we don’t know what to do with it” (Johnson, 2001). This legitimate and important concern is being echoed by police administrators across the country. This article attempts to address the question of how to best collect and examine benchmarks for comparison against pedestrian and traffic stop data.

This article begins by documenting the recent trends nationwide in traffic stop data collection. Six different types of baseline or benchmark measures used in current studies of police citizen contacts are reviewed and compared. Thereafter, the methodology implemented in an ongoing study of traffic stops by the Pennsylvania State Police (PSP) is described in detail. Specifically, four different benchmarks used for the Project on Police-Citizen Contacts (PPCC) in Pennsylvania are outlined, and their strengths and weaknesses are assessed. The sampling, procedures, and techniques used in the PPCC may serve as a guide for data collection efforts in other
large municipal or state agencies. The article concludes with a discussion of how data collection efforts at the local and state levels can improve by experimenting with and comparing benchmark measures.

**TRENDS IN TRAFFIC STOP DATA COLLECTION**

The growth in scholarly research on the topic of police-citizen contacts is a product of several factors, including the increasing legislative and judicial mandates for data collection, departments' proactive responses to the national trend toward data collection, and the Department of Justice recommendation that police departments partner with academics for data collection design and analysis (Ramirez et al., 2000; Strom, Brien, & Smith, 2001). State legislatures have been actively debating racial profiling in the last several years. A 1999 state police agency survey indicated that only two states (North Carolina and Connecticut) had passed legislation that mandated officers to collect race data for all traffic stops (Strom & Durose, 2000). As of 2001, however, 18 states passed legislation addressing racial profiling, most of which also mandated data collection (National Conference of State Legislatures, 2001). Furthermore, many other state police agencies indicated that their state legislatures were considering legislative action on racial profiling (Strom et al., 2001). The trend in most of the existing and pending legislation is toward mandatory, universal, and ongoing data collection efforts (Institute on Race and Poverty, 2001). Other state and local agencies are involved in data collection efforts as a result of Department of Justice investigations or consent decrees (Police Foundation, 2001).

As more police agencies come under legislative or judicial mandate to collect traffic stop data, several state and local law enforcement agencies are taking the initiative to begin data collection efforts on their own. A survey of the 49 state law enforcement agencies in 2001 found that 16 agencies required all of their officers with traffic patrol duties to record drivers’ race and ethnicity for traffic stops, more than doubling the number of state agencies collecting data on all stops in 1999 (Strom et al., 2001). Nearly half of these agencies were collecting data as a result of a self-initiated policy. Another 23 state police agencies required officers to collect information on drivers’ race and ethnicity in more specific circumstances (e.g., arrest, use of force, etc.) (Strom et al., 2001). The majority of the state agencies that collected race data for all stops, as well as five other state agencies with more limited race data collection, made these data available to the public.
The most frequent tool for the dissemination of these data has been the Internet, as the reports of many state and local agencies are available online (Police Foundation, 2001).

Findings from most data collection efforts have indicated that some disparities exist for police stops, citations, searches, and arrests of different racial groups (for a review, see Engel et al., 2002; Walker, 2001). The interpretation of these disparities, however, has varied widely, with some studies concluding that racial prejudice and discrimination are the cause of the disparity and other studies offering more race-neutral explanations. Engel and her colleagues (2002) have suggested that the reason such varied interpretations exist is that "there is no agreement about what constitutes a reliable and valid base rate" (p. 202). Therefore, the continual improvement and comparison of multiple benchmark measures represents one of the most important endeavors in research on traffic stop data and police practices.

BENCHMARK COMPARISONS

Interpretation of data on police-citizen contacts depends on the type of data with which it is compared (Engel et al., 2002). Often referred to as the "baseline," "base rate," "benchmark," or "denominator," most studies examining racial disparities compare police stop data with the expected rate of stops of minorities, assuming no racial discrimination or prejudice. That is, comparisons are made between officers' rates of actions (e.g., stops, citations, searches, arrests) toward minorities to the "expected probability" of these actions toward minorities (Rojek, Rosenfeld, & Decker, in press). Often researchers calculate a "disparity" or "disproportionality" index based on the difference between the actual and expected rates of police action (e.g., Cox, Pease, Miller, & Tyson, 2001, Rojek et al., in press). Most of the initial studies of this kind relied on population estimates to determine "expected probabilities." More recently, social scientists conducting these studies have focused their efforts on improving and refining various benchmark comparisons. Specifically, studies of police-citizen contacts have relied on six primary techniques for gathering benchmark data: census data, observations of roadway usage, official accident data, assessments of traffic-violating behavior, citizen surveys, and internal departmental comparisons. In this section, we describe each of these techniques, provide examples from ongoing or recently completed research, and highlight their relative strengths and limitations.
CENSUS DATA

Estimates of population figures provided by the census are the most widely used benchmark measures for studies of police-citizen contacts. Most of the early studies of racial profiling used racial percentages of the total population as the standard of comparison for traffic stop data (e.g., American Civil Liberties Union, 2000; Cox et al., 2001; Lansdowne, 2000; Spitzer, 1999; Texas Department of Public Safety, 2000; Verniero & Zoubek, 1999). As researchers became more sophisticated with census measures, they created more refined population baselines that better estimates roadway users by using racial percentages of the driving-age population (Cordner, Williams, & Zuniga, 2000; Harris, 1999; Smith & Petrocelli, 2001; Zingraff et al., 2000). Most of these comparisons have been made at state and city levels. Recently, however, researchers have analyzed census benchmark measures at smaller units of analysis, including the precinct (Fagan & Davies, 2000), census tract, and block levels (i.e., Miami-Dade Police Department’s Racial Profiling Study) (Alpert, 2003).

The use of census data benchmarks has been vigorously criticized because population estimates for particular areas may not accurately represent the driving population in those areas (Cox et al., 2001; Walker, 2001; Zingraff et al., 2000). The residents of an area must roughly represent the drivers in that area for census data to provide valid comparisons. However, national transportation and travel surveys data have indicated that racial, ethnic, age, and gender differences exist for driving frequency. For example, analyses using the 1995 Nationwide Personal Transportation Survey (NPTS) data show that Blacks were less likely than both Whites and Hispanics to have a driver’s license. In addition, minorities are 6 times more likely than Whites to use public transit rather than personal use vehicles as their primary means of transportation (Bureau of Transportation Statistics, 1997). Furthermore, reports from the NPTS showed that Blacks were considerably more likely to live in households without a vehicle, compared to Whites (Federal Highway Administration, 1995). Finally, preliminary analyses from a study of several locations in one jurisdiction in Michigan suggest that population statistics overrepresent or underrepresent the non-White driving population anywhere from 7% to more than 300% (Lamberth, 2001, p. 3). Therefore, it is necessary for studies examining differences in rates of police-public contacts during traffic stops to use more precise benchmarks that capture differences in the frequency and patterns of driving behavior by race, ethnicity, and gender.
Rojek et al. (in press) have developed an estimation procedure to address some of the shortcomings of the use of residential populations for benchmarking purposes. Based on concerns from police administrators that population statistics do not accurately reflect driving populations in their districts, Rojek and his colleagues developed a measure of the driving population based on spatial weighting. Using mapping software and spatial statistics, an imputation procedure was developed to better estimate driving populations by giving greater weight to residents, nearby nonresidents, and nonresidents from larger municipalities. The estimated driving populations based on spatial weights differed significantly from straight census population figures and better estimated roadway usage in two of the three locations where it was used. This technique represents a promising use of spatial statistics to better utilize census data for benchmark comparisons.

Observations of Roadway Usage

Another benchmark that has been used compares groups’ rates of being stopped, cited, searched, or arrested to observational estimates of those groups’ driving patterns and frequency. These benchmarks are measured by observing the roadways to record the race and ethnicity of passing motorists for several geographic locations. Lamberth (1994) first used this technique to gather data on the New Jersey Turnpike’s racial and ethnic distribution of drivers as well as in-state and out-of-state vehicle registration. Lamberth randomly selected the dates and times of 18 roadway observation sessions conducted during daylight hours in a 2-week period in June 1993. Observers were located at two stationary observation posts on the New Jersey Turnpike and alternated between the two locations several times a day. Observers recorded the race of the driver and any visible passengers as well as the state of registration for the vehicle. Lamberth noted that the observation teams reported a high degree of confidence in their ability to observe and record the required data. Using a simple White-Black dichotomy, observers were able to identify the drivers’ race in 99.6% of the 42,706 vehicles observed. Drivers that were identified as Black made up 13.5% of all drivers observed, a percentage that is significantly smaller than the 73% of individuals arrested and the 35% of those stopped in a previous 3-year period.

Recently, observation techniques have been used to validate other less costly, less labor-intensive types of benchmarks. For example, the validity of the modified census tool created by Rojek et al. (in press) was examined
by comparing its racial and ethnic estimates with the composition of the driving population as measured by roadway observation. As noted by the researchers, other measures of roadway usage, including observations, “will continue to be necessary to validate the statistical procedures of the sort we have presented and to gather data that is unavailable by any other means” (Rojek et al., in press).

Using roadway observation to gather benchmark data may be particularly useful in areas whose residential and driving populations are unlikely to be the same. Nevertheless, this method does have some important limitations. First, as noted by several scholars, this type of data collection technique is difficult to implement, costly, and time-consuming (Engel et al., 2002; Rojek et al., in press). Second, the reliability and validity of observers’ perceptions of drivers’ characteristics cannot be directly assessed. Third, as described below, direct observation of roadway usage is not an appropriate benchmark if violating behavior differs by drivers’ characteristics.

OFFICIAL ACCIDENT DATA

A recent and innovative approach to assessing roadway usage has been developed by the research team examining police stop data in Metro-Dade, Florida. Alpert, Smith, and Dunham (2003) have considered the use of “not-at-fault driver demographic data from two-vehicle crashes” to assess who is using the roadways. These scholars argued that other fields (e.g., actuarial statisticians and safety engineers) have used automotive accident data for decades to examine relative risks of accidents for particular demographic groups. Alpert et al. (2003) argued that not-at-fault accident victims involved in two-car crashes should approximate the racial composition of drivers in those areas. Obviously, using at-fault accident victims to assess roadway usage could potentially be a biased measure. It is possible that different types of drivers exhibit more or less aggressive driving behaviors that would put them at an increased risk for accidents. Thus, at-fault drivers may not accurately reflect the driving population. The argument for using not-at-fault driver demographics in two-car accidents as an indicator of roadway usage is based on the proposition that drivers who are officially determined to be not-at-fault would not exhibit any types of aggressive (or nonaggressive) driving behaviors that would put them at increased risk for an accident. It is argued that not-at-fault drivers involved in two-car accidents are involved in those accidents randomly and therefore would be a better representation of who is using the roadways.
The use of accident data for the creation of benchmark comparisons is an interesting yet untested methodology. The utility of this approach will likely hinge on the availability of this type of data along with the ability to disaggregate the data into specific locations, times, and days of the week that can be matched directly with traffic stop data. Note, however, that at these levels of aggregation, there is unlikely to be a sufficient sample size to make meaningful projections to the driving population. Furthermore, findings from accident data should be compared to observation data. If similar trends emerge, the use of accident data may represent a more cost-effective way to capture this information at larger levels of aggregation.

ASSESSMENTS OF TRAFFIC VIOLATING BEHAVIOR

Examining who is using the roadways does not address the larger question of who is more likely to be stopped by police for traffic violating behaviors. Social scientists and police officials are exploring the possibility that particular types of citizens (e.g., young African American males) may drive more aggressively and therefore may be more likely to violate traffic laws and/or commit more serious violations (Lange, Blackman, & Johnson, 2001; MacDonald, 2002; Smith et al., 2000). Neither the population estimates nor the roadway usage surveys described above, however, examine the possible differential offending patterns of White and non-White motorists. Although a national survey by the National Highway Traffic Safety Administration observed both gender and age differences in driving frequency, speeding, and other unsafe driving behavior (Boyle, Dienstfrey, & Sothoron, 1998), few studies have examined race- and/or ethnicity-based differences in law-violating driving behavior. Note, however, that some studies have shown that African American and Hispanic drivers have slightly higher rates of involvement in both accidents and alcohol-related accidents compared to White drivers (Royal, 2000; Voas, Tippetts, & Fisher, 2000).

The possibility that groups differ in their driving patterns or in their frequency and degree of law-violating driving behavior is an important race-neutral explanation of disparity that has been considered. Lamberth (1996) initiated the first efforts to establish a baseline of law-violating driving behavior in separate studies in New Jersey and Maryland. In what he described as the “carousel method,” observers rode in cars that were driving with the cruise control set at 5 miles per hour over the speed limit in New Jersey and exactly at the speed limit in Maryland. The observers counted the
number of cars that passed them (i.e., traffic violators) and the number of cars that they passed (i.e., nonviolators) while recording the race of the driver in each car. Using this technique, Lamberth reported that 98% and 93% of drivers in New Jersey and Maryland, respectively, were considered traffic violators; however, White and Black drivers drove indistinguishably.

A major limitation of Lamberth’s initial methodology (1994, 1996) to examine differences in speeding was the inability to determine how far above the speed limit motorists were traveling. Most police organizations have either an informal or a formal policy regarding the amount above the speed limit at which citations are issued. Thus, the degree of law-violating behavior measured in Lamberth’s studies did not capture the seriousness of motorists’ driving infractions and, as a result, their real risk of being stopped by police.

Since Lamberth’s initial studies (1994, 1996), other researchers have improved on this data collection methodology. Research efforts in North Carolina have expanded on Lamberth’s technique to better estimate the degree to which drivers violate the speed limit (Smith et al., 2000). In this study, groups of observers used stopwatches to measure the time that it took vehicles to pass the distance from the rear bumper to the front bumper of the observer’s vehicle (traveling at a set speed). Although this method systematically underestimated the speed of the passing vehicle, researchers were able to systematically correct for this underestimation so that estimated speeds corresponded to actual speeds. In addition to the speed of passing vehicles, observers also recorded information about the vehicle and its occupants (e.g., drivers’ race, gender, approximate age, vehicle color, state of license plate, type of vehicle).

Although the research conducted in North Carolina addresses the important issue of differences in the severity of law-violating behavior (e.g., speeding) by gender and race, it still does not fully examine drivers’ real risks of being stopped by police for speeding. As acknowledged by these researchers, motorists differ in their levels of “speeding savvy” (Smith et al., 2000). That is, some motorists are better than others at speeding without being detected by police officers (e.g., motorists that travel close to other vehicles or near tractor trailers, have RADAR detectors, routinely slow down with traffic, etc.). Therefore, citizens’ risks of being stopped for speeding are not fully captured through methods that strictly examine differences in the severity of speeding behavior.

A more recent examination of traffic-violating behavior and potential detections of speeding savvy was conducted in the Speed Violation Survey
of the New Jersey Turnpike (Lange et al., 2001). This study used RADAR and high-speed photography at 14 different locations along the 148-mile turnpike to identify the race, ethnicity, gender, and speeding behavior of drivers on the roadway. Each location yielded approximately 48 hours of data collection during a 3-month period in 2001, which was varied by weekend and weekday. The researchers operationalized speeding as driving at least 15 miles per hour over the posted speed limit.

Lange et al. (2001) used a panel of three trained observers working independently to identify the drivers’ race, ethnicity, gender, and age from the photographs. Cases with at least two identical ratings were treated as conclusive, and others, unclassifiable. Out of 38,747 photographs, two of the three coders agreed on the race/ethnicity of 26,334 drivers, or 68% of the photographs. Findings indicated that of the drivers whose race/ethnicity was identified, Black drivers were 64% more likely than White drivers of similar age and sex to exceed the 65-mph speed limit by 15 mph or more. Drivers classified as “other” were 18% more likely than Whites to be speeding in the 65-mph zone. At the 55-mph speed limit, no statistically significant differences between Blacks and Whites were found. However, Hispanic drivers and drivers classified as “other” were less likely than Whites to be speeders. Overall, the vast majority of drivers were not found to be driving faster than 15 mph over the posted speed limit, which suggests that the criterion used for speeding in this study may have been too high.

Another limitation of this study acknowledged by the authors was that the number of Hispanic drivers might have been underestimated. Only 4.8% of drivers were identified as Hispanic, compared to 14.2% of drivers self-identified as Hispanic in the companion survey of drivers also conducted on the New Jersey Turnpike (Farmer, 2001; Lange et al., 2001). Similar differences between the Black and White populations of the two surveys were not found. Furthermore, aggregate comparisons of White and non-White groups suggest that evaluators may have incorrectly classified Hispanic drivers as White.

This study has been criticized by the Justice Department and representatives of the NAACP for the use of photographs and the potential bias associated with being unable to use nearly a third of the photographs collected (Fears, 2002; Seper, 2002). Although the study reported a statistically significant overrepresentation of unusable data among speeders, Lange et al. (2001) found no evidence to indicate that the unusable data differed significantly by drivers’ race. The authors suggested that unusable data was
primarily due to technical problems associated with the positioning of cameras that produced glare and shadows on the windows of passing cars.

A general limitation of observational surveys that rely on the use of RADAR for speed detection may be that its use slows down the speed of passing traffic. However, it is likely that the effect of surveyors' or observers' use of RADAR on traffic is similar to the effect of officers' use on driver behavior (Lange et al., 2001). Another limitation of all benchmark comparisons that try to assess traffic-violating behavior is that it is difficult to reliably measure all driving behaviors that violate the law (Lange et al., 2001). Surveys of speeding behavior, however, might be appropriate benchmarks for agencies that initiate the majority of their traffic stops for speeding rather than other driving violations. Observational research conducted in 16 locations within the Metro-Dade area as part of the Miami-Dade Police Department's Racial Profiling Study addressed this limitation by assessing three different types of traffic violating behavior: speeding, running red lights, and not coming to complete stops at stop signs (Alpert, 2003).

CITIZEN SURVEYS

Citizen surveys that examine issues of racial and ethnic profiling have generally taken two forms: (a) surveys designed to capture information regarding interactions with police and (b) surveys designed to create baselines for comparison with traffic stop data. The first type of citizen survey asks respondents about their contacts with police (rather than relying on administrative police data). Surveys may be administered to a random or stratified sample of citizens (e.g., the Police-Citizen Contact Survey 1999), or to a sample of citizens who have received citations during recent traffic stops (e.g., Wright, Tomaskovic-Devey, & Zingraff, 2000). Similar to police-stop data, this type of information collected through citizen surveys must be compared to some type of benchmark measure.

In contrast, surveys that examine roadway usage and travel patterns more generally can be used to create benchmarks for comparison with traffic stop data. For example, citizen surveys of roadway usage include national surveys that inquire about respondents' use of motor vehicles and other types of transportation (e.g., the National Personal Transportation Survey 1995). As described by Pickrell and Schimek (1998), data for the National Personal Transportation Survey can be used to derive estimates of drivers' annual mileage to examine how often and when particular types of people...
are using the roadways. As previously noted, straight population figures may be inappropriate comparisons to traffic and pedestrian stops. The level of inaccuracy in these baseline measures, however, is unknown. Using traffic surveys would allow researchers to compare racial, ethnic, and gender disparities in traffic stops and police behavior by better approximating citizens’ driving patterns and frequency.⁵

An additional option is to gather demographic information from respondents as they exit tollbooths or other roadway areas. For example, a research team from the Public Services Research Institute conducted interviews with drivers in New Jersey along three different segments of the turnpike, which corresponded to the locations of the New Jersey State Police stations along the highway. The research team randomly contacted 4,656 drivers as they exited the turnpike at various locations during a 2.5-week period in 2000, which varied by time of day and day of week. Eighty-seven percent of the drivers agreed to participate in the interviews and identified their own racial and/or ethnic categories (Farmer, 2001).

Although the findings about the racial and ethnic makeup of the New Jersey Turnpike’s driving population are useful for establishing more accurate baselines of driving patterns and frequency, the results of the survey have implications for other methods of measuring roadway usage. This survey of drivers found that the turnpike-driving population was far from uniform, noting differences in the racial and ethnic composition of drivers by day of week, time of day, geographical location, and even the direction of travel on the roadway (Farmer, 2001). Thus, roadside surveys and observational studies must be conducted in several locations over an extended period of time to best estimate the driving population.

The primary strength of using citizen surveys to establish comparative benchmarks is that individuals are able to self-select their racial and/or ethnic identity, whereas observational studies are less able to reliably assess racial and/or ethnic group membership. On the other hand, a significant weakness of citizen surveys is that they are less equipped to measure traffic-violating behavior because relying on citizens to divulge their illegal driving behaviors may introduce socially desirable responses. For example, a survey of drivers that had been issued citations in the past 6 months by the North Carolina Highway Patrol indicated that 30% of drivers did not accurately divulge that they had received a driving citation (Wright et al., 2000). It also appears that reliability of responses differed significantly by drivers’ race as White speeding violators were more likely to indicate that they had
received a citation compared to Black speeding violators (Wright et al., 2000).

**INTERNAL DEPARTMENTAL COMPARISONS**

An alternative benchmark used by some agencies is comparing officers' rates to other officers. Rather than creating external benchmarks, internal departmental comparisons allow for the use of police-citizen contact data without additional external data. Essentially, officers' rates of stopping, citing, searching, and arresting minority citizens are compared to other officers working in the same or similar assignments, areas, and/or shifts. These comparisons are often part of larger management tools, called “early intervention” or “early warning” (EW) systems, which are used by police departments to identify problem officers (Walker, 2001). Samuel Walker (2001, p. 84) has argued that EW systems are a “promising but not fully proven” tool for achieving police accountability that can easily be used to examine rates of police-citizen contacts. These systems are mostly used by agencies that are under court or legislative mandate to collect police-citizen contact data.

Walker (2001) has acknowledged two general limitations of using the EW approach to examine traffic and pedestrian stop data. First, he suggested that EW systems are not effective in cases in which “an entire agency is engaging in racial or ethnic discrimination” since “the behavior of all officers will be roughly the same” (Walker, 2001, p. 87). However, the use of internal departmental comparisons to examine this type of data is even more limited than Walker has suggested. Although it is true that an EW system could not identify an entire department engaging in discriminatory practices, the same argument could be made for districts, shifts, or beat assignments within a department. Without comparisons to some other criteria, it would be very difficult to determine the validity of using police activity measures as the denominator, rather than the numerator, when examining racial disparities. As noted by Fagan (2002), internal activity measures are only one part of a more complex assessment of whether police actions are racially disparate and whether these disparities rise to the threshold of a “profile.” To address these questions, additional baseline data are required from which the predictive dimensions of police actions can be observed. (p. 8)
A second limitation acknowledged by Walker (2001) is that EW systems require that individual officers be identified on police-citizen contact data collection forms. The possibility of civil and criminal liability for individual officers, coupled with the powerful opposition of police unions, has led many police departments to preclude individual identifiers. Thus, internal comparisons are not feasible for most agencies; the data available simply cannot establish officer comparisons. Although Walker (2001, pp. 88-89) effectively argued for including individual identifiers in future collection of traffic stop data, most police departments and unions remain understandably reluctant to collect this information given the political and legal climate and the media attention surrounding racial profiling. Therefore, at this time, the use of internal comparisons remains a very limited method for examining trends in traffic and pedestrian stops.

Each of the benchmark measures reviewed—census data, observations of roadway usage, official accident data, assessments of traffic-violating behavior, citizen surveys, and internal departmental comparisons—have strengths and weaknesses. The feasibility and accuracy of their collection will depend greatly on the type of police department, geographic location, and resources available. Given the inherent biases of each measure, our recommendation is to use multiple measures to ensure direct comparisons between benchmarks and police-citizen contact data. The PPCC for the Pennsylvania State Police Department recently took this approach, and their techniques used for creating benchmark measures are described below.

**PROJECT ON POLICE-CITIZEN CONTACTS**

The PPCC was implemented in January 2002, and data collection began department-wide in May 2002. Although not under legislative or court mandate to collect traffic stop data, officials from the PSP decided to engage in a large-scale data collection effort examining police-citizen contacts during all member-initiated traffic stops with citizens across the Commonwealth. The PSP contracted with a research team from The Pennsylvania State University to develop the data collection instrument and provide an independent assessment of the data.

A committee of PSP administrators, in partnership with the principal investigator, drafted the original data collection instrument for use by troopers. This data collection form was pilot tested in one station for 4 weeks during all member-initiated traffic stops. Based on insights gleaned from the
pilot test, the form was altered slightly, officers were trained, and the data collection effort was expanded department-wide for 1 month. The full 12-month data collection effort was implemented after feedback was delivered to field supervisors regarding troopers’ compliance with the data collection effort and nonrandom errors on the forms. Data collected includes information regarding (a) the stop (e.g., date/time, location, type of roadway, reasons for the stop, and the duration of the stop), (b) the driver (e.g., gender, age, race/ethnicity, zip code of residency), (c) the vehicle (e.g., state of registration, number of passengers), (d) the outcome of the stop (e.g., citation, written warning, arrest, search, property seized during the search), and (e) identification information (e.g., location of the stop—county and municipality and the trooper’s station and employee identification).

To better examine and interpret the police-citizen contact data, the academic research team collected several different benchmark comparisons. Specifically, the police-citizen contact data collected by Troopers will be compared to four related benchmark measures: (a) census data of residential populations where the traffic stop occurred, (b) census data of residential populations where the motorist resides, (c) observations of roadway usage, and (d) surveys of traffic-violating behavior. Each of these techniques is described in detail below, following a brief description of the PSP.

**THE PENNSYLVANIA STATE POLICE**

The PSP is the oldest (established in 1905) and largest full-service state police agency in the country. The PSP serves the Commonwealth of Pennsylvania, with an estimated population of 12,281,054 in 2000, including 15% minorities, 5.6% unemployed, 10.6% below the poverty level, and 6.2% female-headed households with children (Pennsylvania Department of Labor and Industry, 2002; U.S. Census Bureau, 2002). The violent crime rate in 2000 was 402 per 100,000 residents, below the national rate of 506 per 100,000 (FBI, 20XX; Pennsylvania Uniform Crime Reporting System, 2000). In 2002-2003, the department employed 4,118 sworn officers, 4.2% of whom were female, 10.9% minority, 29.7% with a 4-year college degree, and who had an average of 12.1 years of service.

The PSP is divided into 14 different bureaus (e.g., Bureau of Patrol, Bureau of Professional Responsibility, Bureau of Research and Development, etc.) under the direct supervision of the commissioner and three deputy commissioners (all of whom are appointed by the governor). The operational jurisdiction of the PSP is geographically divided into five areas,
16 troops, and 90 stations, with a ranking commander at each level (Pennsylvania State Police, 2000).

**CENSUS DATA FOR TRAFFIC STOP LOCATION AND MOTORISTS’ RESIDENCE**

Most traffic and pedestrian stop studies have relied at least partially on population statistics for benchmark comparisons. Often, statewide population statistics or statewide driving-age population statistics are used. However, sole use of state-level populations is inappropriate because of the geographic clustering of racial and ethnic populations. Pennsylvania is no exception: for example, compare the overall statewide percentage of Blacks (10.5%) to the population of Blacks within the cities of Harrisburg (57.5%) and Scranton (3.6%) (U.S. Census Bureau, 2002). Even county- or city-level estimates may be inappropriate comparisons in areas where commuters, tourists, or long-distance travelers are the primary users of interstates and highways. Ultimately, examining traffic stop data at lower levels of aggregation is necessary. Although data collection efforts in local municipalities can focus at the census tract or even block level (Alpert, 2003), data collection efforts in state police organizations must focus on larger units of analysis. The PPCC examines police behavior at the county, municipality, and residential zip codes levels.

In Pennsylvania, there are 67 counties, 2,567 municipalities, and 2,111 residential zip codes. Municipalities are political subdivisions of counties and incorporate cities, boroughs, towns, and townships (Census Bureau, 2002). Zip codes primarily identify regions and metropolitan areas within the United States for the purposes of mail distribution but do not necessarily conform to other jurisdictional boundaries (United States Postal Service, 2003). For each member-initiated traffic stop by PSP troopers, the municipality and county of the stop is recorded, along with drivers’ residential zip codes. Using population statistics for each municipality, a disproportionality index was developed to compare rates of minority stops to rates of minorities in the population (Cox et al., 2001; Rojek et al., in press). The collection of drivers’ residential zip codes, however, will allow for the inclusion of data only for those drivers who reside in each particular municipality. Using residential zip codes provides a better estimate of traffic stops of nonresidents than does in-state/out-of-state registration comparisons. The collection of residential zip codes will also indicate if the use of municipality- and county-level population statistics is appropriate.
Analyses to be conducted on the completed data set will use mapping and spatial statistics to examine whether adjacency and other measures of proximity can better estimate differences between location of residence and location of the stop (e.g., Rojek et al., in press).

ROADWAY USAGE AND TRAFFIC-VIOLATING BEHAVIOR SURVEYS

There are a number of important reasons to consider both the roadway usage and traffic-violating behaviors of different groups of motorists. The PSP contracted for an independent survey of roadway usage and traffic-violating behavior given the importance of supplementing census-level data that may not provide accurate comparisons to traffic stop data. Due to financial and administrative constraints, however, it was impractical to conduct observations of roadway usage and traffic violating behavior statewide. The mere size of most states would make a statewide observation study cost-prohibitive. Thus, the following description of the sampling and procedures implemented for the roadway usage and traffic-violating surveys for the PSP may serve as a guide for social scientists and police administrators examining larger geographic areas for comparisons to traffic stop data.

County Sampling Procedures and Roadway Location Selections

A decision was made to sample counties for inclusion in the roadway usage and traffic-violating behavior surveys, rather than to observe traffic patterns in all 67 counties across Pennsylvania. Given the directed nature and specific issues surrounding the need to collect benchmark information, seven characteristics of counties were identified to assist in the sampling selection process, instead of conducting a simple random sample of counties. The research strategy was to identify characteristics of counties related to three specific concerns: (a) general roadway usage, (b) roadway usage by minorities in particular, and (c) roadway usage by drivers that may not directly reflect surrounding residential populations. The variables measured to address these concerns are described below and included in Table 1.

The first concern, general roadway usage, was assessed through the measurement of three variables: (a) the total population, (b) the number of interstate miles, and (c) the total number of roadway miles for each county.
Population size was measured using 2000 census data, and the interstate and roadway miles were measured using data collected by the Pennsylvania Department of Transportation (2001).

The second concern, roadway usage by minorities, was estimated through proxy measures capturing the population of Blacks and Hispanics in each county from the 2000 Census. Although studies have shown that minorities compared to Whites are more likely to use public transportation (Bureau of Transportation Statistics, 1997), more likely to live in a household without a vehicle (Federal Highway Administration, 1995), and less likely to have a driver’s license (Langan, Greenfeld, Smith, Durose, & Levin, 2001), it is still reasonable to expect that larger concentrations of minorities would lead to more roadway usage, compared to areas with lower concentrations of minority populations. Therefore, the population of Blacks and Hispanics in a county is used as crude indicators of likely roadway usage by minorities.

The third and final concern—roadway usage by drivers that may not directly reflect surrounding residential population—was estimated through two dichotomous variables: (a) the presence of tourist attractions, colleges and universities, or historical sites, and (b) the presence of seasonal attractions (e.g., amusement parks, water parks, ski resorts, etc.). The intent was to identify particular counties where the traffic patterns may not accurately reflect full-time residency captured in census data. Furthermore, the seasonal attractions variable was designed to identify counties where traffic volume and the racial composition of drivers may fluctuate during particular months of the year.

### TABLE 1. Variables Used to Select Counties for Observation

<table>
<thead>
<tr>
<th>Variable</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Factor Loading</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total population</td>
<td>4,946</td>
<td>517,550</td>
<td>183299.31</td>
<td>265442.82</td>
<td>0.91</td>
</tr>
<tr>
<td>Percentage Black</td>
<td>0.10</td>
<td>43.20</td>
<td>3.39</td>
<td>5.88</td>
<td>0.75</td>
</tr>
<tr>
<td>Percentage Hispanic</td>
<td>0.30</td>
<td>10.20</td>
<td>1.86</td>
<td>2.25</td>
<td>0.56</td>
</tr>
<tr>
<td>Interstate miles</td>
<td>0</td>
<td>93.80</td>
<td>26.32</td>
<td>24.06</td>
<td>0.66</td>
</tr>
<tr>
<td>All roadway miles</td>
<td>312.40</td>
<td>5670.8</td>
<td>1785.71</td>
<td>966.91</td>
<td>0.75</td>
</tr>
<tr>
<td>Tourist attractions, historical sites, colleges/universities</td>
<td>0</td>
<td>0.70</td>
<td>0.46</td>
<td>0.49</td>
<td></td>
</tr>
<tr>
<td>Amusement parks/seasonal attractions</td>
<td>0</td>
<td>0.13</td>
<td>0.34</td>
<td>0.61</td>
<td></td>
</tr>
</tbody>
</table>

Note: n = 67 counties.
These seven characteristics—population total, percentage Black, percentage Hispanic, number of interstate miles, number of total roadway miles, presence of tourist attractions/historical sites/colleges and universities, and presence of amusement parks/seasonal attractions—were analyzed using principal components factor analysis to identify any underlying latent constructs. The factor analysis revealed a factor with an eigenvalue greater than 1 (eigenvalue = 3.31), which explained 43.7% of the variance.

The specific factor loadings for each variable are displayed in Table 1. Individual factor scores were generated for each county, and the counties were ranked from high to low based on these scores. Thus, essentially, the counties were ranked on their potential volume of traffic, possible minority roadway usage, and possible travel patterns that do not match residential populations.

The ranked 67 counties were then divided into four groups ranging from high to low levels of the above-listed variables, and 20 counties were selected for observation. Of the 20 counties selected, 55% (11 counties) were from the high group, 20% (4 counties) were from the medium group, 15% (3 counties) were from the medium/low group, and 10% (2 counties) were from the low group. This selection process represents an aggressive oversampling of the “high” group to better examine the counties where there is likely to be more traffic, more minority roadway usage, and traffic patterns that may not reflect residential populations. The final selection of counties from within these four groups was also based on the amount of departmental activity within those counties and their geographic location.

Once the counties were identified, specific roadways to be observed within those counties were selected. The research team deferred to the expertise of PSP station commanders (lieutenants) working in those counties to select the initial locations for data collection, based on a set of predetermined criteria. Specifically, station commanders were asked to identify multiple locations for observation based on the following criteria: (a) locations with larger volumes of traffic, (b) locations representative of major travel patterns in the county, (c) locations that were good RADAR sites (i.e., areas that generate the most citations from troopers), (d) locations where observers would be able to use RADAR and still observe drivers (e.g., not on an overpass or too far from the roadway), and (e) locations that were safe for observers to be parked off the roadway. As the project progressed, the locations identified by station commanders as generating the most citations were compared to the number of citations generated in those municipalities during the previous year and the percentage of stops as recorded on the
current police-citizen contact forms. The majority of locations selected did accurately reflect statistical comparisons of citations and stops.¹¹

**Procedures, Training, and Data Collection**

Trained undergraduate research assistants were responsible for the collection of data for both observations of roadway usage and traffic-violating behavior surveys. Observers participated in two 8-hour days of training. The first training session was conducted by PSP officials at the State Police Training Academy. Observers received 4 hours of classroom instruction on the philosophy, use, and limitations of RADAR. They were then escorted to an interstate to practice using the RADAR units under the instruction of PSP troopers. The second day of training consisted of 4 hours of classroom instruction to document the specific procedures for observation and collection of data, followed by roadside training using the actual data collection instrument.

During data collection, two observers parked in a personal vehicle on the side of the roadway collected information about the characteristics of passing motorists, information about the vehicle, and the speed of the passing vehicle via RADAR. Observers recorded the race of the driver, other relevant driver information, and vehicle information to the best of their abilities. Observers were trained that they must separately measure (and agree) on the race/ethnicity of the driver using detailed categories (e.g., White, Black, Hispanic, Asian, etc.), which corresponded directly to the data collected by troopers during traffic stops. If observers could not determine the ethnicity of the driver but could determine race, a dichotomous Caucasian/non-Caucasian category was used. If both observers did not agree on the Caucasian/non-Caucasian category or the race of the driver simply could not be determined, the data were coded as missing. Observers were scheduled for data collection during daylight hours and in weather conditions that allowed optimal visibility.

As of this writing, the PPCC data collection effort is in progress. Within each selected county, research assistants will conduct a total of 10 days of observation (7 to 8 hours per day) or approximately 75 hours of observation in each county, for a total of 1,500 hours of observation. Days of the week, times of day, and months of the year will vary to allow for day, time, and seasonal variation in traffic patterns. The information will be compiled to generate average information regarding drivers' road usage and law-violating behavior in the selected counties and municipalities. Results from the training period conducted prior to observers' participation in the study and from
the first 6 months of data collection indicate that observers are able to readily capture the race and other characteristics of passing motorists.

STRENGTHS AND LIMITATIONS OF METHODOLOGY

As with any social science research, there are several limitations of this methodology. It is important to note, however, that the study of racial and ethnic profiling by police departments is a continually improving scientific process. Despite the difficulties of studying police-citizen contacts and the political pitfalls associated with research designed to investigate racial and ethnic disparities, studies in this area have advanced quickly. Yet the limitations of research designs currently used (including the methodology outlined above) are numerous and must be appropriately recognized to help this field of study advance.

One of the most salient criticisms of the data collection proposed above involves the subjectivity of identifying drivers’ race and ethnicity. The issues regarding the validity and reliability of drivers’ race and ethnicity measures are slightly different for data collected by the police during traffic stops compared to data collected by observers to assess roadway usage and speeding. For data collected directly by police, the reliability and validity of citizens’ race involves two related concerns. First, police may be reluctant to indicate drivers’ race or may simply report inaccurately. Second, officers may “disengage” or initiate fewer traffic stops overall. Both of these behaviors represent an effort by officers to protect themselves from criticism, departmental discipline, and potential lawsuits. From the officers’ perspective, this is a reasonable response to data collection efforts specifically designed to identify officers who “racially profile.”

Unfortunately, the validity of the data collected by police often cannot be directly assessed. There are several ways, however, to increase validity and reliability of this type of data collection. First, researchers and police officials should consider protecting officers from internal discipline and potential civil and criminal liability based on the data collection effort. Providing confidentiality to officers involved in data collection is controversial. Citizen groups and police administrators often believe officers should be individually identified and held accountable for their actions. This is the direct philosophy behind EW systems. In contrast, police unions and department attorneys often advocate confidentiality for officers due to the misunderstanding and misuse of police-citizen contact data. The trade-off that police administrators and researchers must consider is often between
more reliable data that can ultimately be used for training purposes versus less reliable data that can better identify problem officers.

A second way to increase validity is to crosscheck the data with other measures to assess levels of nonvalid data and officer disengagement. For example, in states where drivers' race is collected by the Department of Transportation (or Department of Motor Vehicles), drivers who are issued citations can be checked through the use of these records. The research team analyzing data for Metro-Dade, Florida, have used license photographs to determine the accuracy of data collected by police officers (Alpert, 2003). More aggressive cross-validation has been used by the North Carolina research team, where citizens who were issued citations were actually contacted and administered a telephone survey (Wright et al., 2000). Scholars are exploring how best to examine the rate of officer disengagement; one technique is to compare levels of activities that are captured on departmental forms other than the traffic stop form (e.g., citations, arrests, searches, etc.) before and after the implementation of traffic stop data collection.

Concerns about the reliability and validity of observers' measurement of drivers' race and ethnicity represents additional limitations of the data collection effort. It is possible that observers are not accurately recording drivers' characteristics. To address this possibility in the Pennsylvania data collection effort, observers were trained that unless they both agreed on a driver's characteristics, they must record the information as missing data. Despite advances in procedures and training, the accuracy of observers' classifications will always be questioned. It is also important to note, however, that unlike the use of photographs, the use of observers allows researchers to collect data in conditions that are somewhat similar to what officers may actually experience. Allegations of racial and ethnic profiling assume that officers can determine drivers' race/ethnicity prior to the initial traffic stop. This is a particularly important issue for state police agencies that make a majority of their stops for speeding. How often drivers' characteristics can be determined in stationary locations using RADAR is an empirical question that has not been addressed. As noted above, training sessions conducted prior to observers' participation in the study indicated that observers can determine the driver's race in good weather, during daylight hours, and when RADAR is conducted in locations with clear visibility to the roadway. Therefore, if the goal of the research is to determine drivers' risk of being stopped for speeding, observers using RADAR in stationary vehicles may be a stronger method than observers in moving vehicles or strategically placed video cameras.
One of the primary reasons to measure this type of benchmark is to capture drivers' demographic information. Accordingly, observers were trained to separately look for and then confer about drivers' race. This objective clearly requires that the observation locations selected be suitable for both RADAR use and the identification of drivers' characteristics. PSP troopers, however, want to detect traffic violations. They are not restricted to RADAR use in these types of locations and often conduct speed enforcement in locations that do not allow for troopers to observe drivers' characteristics prior to the stop. For example, PSP troopers routinely position their vehicles in locations where motorists cannot observe them until after they drive by, conduct RADAR enforcement in teams where one trooper identifies the speeding vehicle and a second trooper stops that vehicle, and use helicopter enforcement. Each of these techniques does not allow the officer identifying the speeding vehicle to simultaneously observe the driver's characteristics. Therefore, it is unlikely that troopers would be able to identify a similarly high percentage of drivers' characteristics prior to the stop.

Another limitation of the proposed research design is that sampling of particular counties for observation does not allow for the comparison of all Pennsylvania counties and municipalities. The costs to do so would be prohibitive for most state and local agencies. Therefore, the sampling design chosen is the best alternative. Some of the comparisons of the traffic stop data will ultimately be made against census data. The collection of drivers' zip codes, however, will allow for more accurate census data comparisons examining only those drivers who reside in the specified locations. In addition, the use of multiple benchmarks collected for 20 of the 67 counties will allow for comparisons between census population figures and actual roadway usage. Given the diversity of Pennsylvania counties, we will be able to determine whether census data is more accurate for particular types of counties (e.g., counties with lower percentages of minorities) and less accurate for others (e.g., counties with seasonal attractions, larger residential populations of minorities, etc.).

**DISCUSSION**

Collectively, the methodology used to examine disparities in traffic and pedestrian stops is not particularly strong. Other areas of study in criminal justice are far more advanced methodologically and statistically. The techniques currently being used by several researchers, however, represent a vast improvement over earlier studies, and the field appears to be advancing
despite the legal and political issues that exist for both researchers and police officials. The growing sophistication in methodological techniques for measuring benchmarks represents a significant contribution to this developing body of research. Researchers and police practitioners have created several different benchmark measures for comparisons with traffic stop data including census data, observations of roadway usage, official accident data, assessments of traffic-violating behavior, citizen surveys, and internal department comparisons. Research conducted for the PSP uses several of these benchmark measures in an effort to make more accurate comparisons to police stop data.

The creation of benchmarks is particularly difficult for police departments with large or diverse geographic areas. Many of the benchmarks described are cost-prohibitive for police agencies with large jurisdictions. However, the sampling technique used for the collection of observation and traffic-violating behavior for the PPCC illustrates a promising solution for larger state agencies. In addition, the PPCC addresses the need to consider multiple issues associated with traffic patterns. For example, in the selection of sample counties for observation, three traffic-related issues were considered: (a) general roadway usage, (b) roadway usage by minorities in particular, and (c) roadway usage by drivers that may not directly reflect surrounding residential populations. Comparisons of different benchmarks will allow future research to better examine and select the appropriate measures for their particular needs.

Some current studies are also examining new issues associated with the creation of benchmark measures. They consider whether traffic-violating behavior varies by citizens’ characteristics (Lange et al., 2001; Smith et al., 2000). This represents a vast improvement over Lamberth’s (1994, 1996) initial studies that documented only whether there were differences in speeding by race and not the actual severity of the offenses. The use of spatial statistics and mapping procedures also represents a new, more accurate use of census data (Rojek et al., in press). Varying levels of aggregation, coupled with multiple measures of proximity and adjacency are among the most sophisticated changes in the use of census data.

Researchers and police practitioners should consider the simultaneous use of multiple benchmarks for comparisons with traffic and pedestrian stop data. Several research projects currently underway or nearing completion will allow for various comparisons of different benchmarks that will guide future research. A major criticism of the studies examining “racial profiling” is that they do not consider how or why officers would
make decisions based on citizens’ race (Engel et al., 2002). New research findings and the continued advancement in the measurement of benchmarks will allow researchers to better test theories of officer behavior that are currently deficient in research examining race-based differences in police-citizen contacts.

NOTES

1. Specifically, African Americans represent 11.8% of all households in the United States but account for 35.1% of the households without a vehicle.

2. For example, state law in Pennsylvania requires that motorists be traveling at more than 6 miles per hour above the posted speed limit to be issued a citation (Pennsylvania Motor Vehicle Code, 75 Pa. C.S. § 3368).

3. The study’s findings did not differ substantially when analyses were replicated using only data that all three coders agreed on.

4. In addition to the racial differences in speeding found in the 65 mph zone, people coded as younger than 45 were 3 times more likely to speed than those older than 45, controlling for race and gender, and men were 20% more likely to speed than women, controlling for race and age. Although the age difference persists in the 55-mph zone, the gender difference disappears.

5. One of the limitations of using national citizen surveys of roadway usage to estimate baselines is that they only provide meaningful comparisons to nationwide police-contact data (e.g., the Police-Citizen Contact Survey). Similar to census data at higher levels of aggregation, national survey data of roadway usage are inappropriate for comparisons to stop data for individual municipal and state agencies. However, conducting citizen travel surveys at the local and state levels might be a viable option that researchers and police administrators should consider.

6. Although the California Highway Patrol has more sworn personnel (approximately 6,700 officers), they are not a full-service police organization (Helmick, 2000).

7. Factor analysis is a statistical technique that, in effect, reduces multiple variables to determine an underlying dimension, or factor, that exists among them. In the case of the variables listed above, each of these variables is highly correlated with the others. Together, the variables represent an underlying dimension or construct. This underlying dimension could be thought of as something that measures larger volumes of travel by minorities or travel patterns that may not match residential populations. For details regarding the use of factor analysis, see Kim and Mueller (1978).

8. In addition, a second factor was extracted with an eigenvalue slightly greater than 1 (eigenvalue = 1.12). However, this factor only explained 16% of the variance, and none of the factor loadings for individual variables were greater than .50. This factor was statistically weak and uninterpretable due to the small factor loadings. As a result, the factor analysis was interpreted as having only one significant underlying factor. The sampling procedures, therefore, were based on the factor scores generated from the main factor. The standardized factor scores for each county are available from the authors on request.
9. The 20 counties that were selected are Allegheny, Bucks, Centre, Chester, Columbia, Dauphin, Delaware, Erie, Franklin, Indiana, Juniata, Lackawanna, Lehigh, McKean, Mercer, Montgomery, Tioga, Washington, Westmoreland, and York.

10. The final selection of counties from the four categories determined by factor analysis was based on input from Pennsylvania State Police (PSP) administrators and the research team. Special consideration was given to the specific activities of the department. For example, some counties were not selected (e.g., Philadelphia county) because PSP has limited jurisdiction in those areas, although other counties were selected because of higher PSP activities. In addition, consideration was given to geographic location in an effort to more effectively cover the entire state and all major interstates. More specific details about the sampling and selection procedures are available from the authors on request.

11. When analyses of the citation and stop data indicated that particular municipalities generating larger numbers of citations and traffic stops were not identified as locations for observation by station commanders, station commanders were specifically asked to select locations for observation in those municipalities.

12. For the data collection effort with the PSP, confidentiality has been contractually promised to each officer. Although officers' employee numbers are initially reported on the data collection forms, the research team is required to strip this information from all data files after officers' demographic information has been successfully merged with the contact data. Through the procedures included in the contract and approved by the University Institutional Review Board, PSP legal team, and PSP union officials, individual officers cannot be identified in data analyses. Each officer was advised of this confidentiality agreement by the principal investigator in a training video.

13. This of course, does not address the possibility that officers racially profile after the stop is made. It is possible that officers are more likely to search, cite, and arrest minorities once they are stopped.

14. The research team has also learned what officers have known all along—that the initial decision to stop a car for a speeding infraction cannot be based on characteristics of the driver alone. Observers (and officers) are trained to identify a car and determine the speed of that car. It is only after a vehicle's speed has been determined and it passes the stationary vehicle using RADAR that drivers' characteristics can be determined. Of course, officers may make decisions to stop vehicles based on this information, but drivers must be violating the law first. That is, for speeding infractions, drivers' race/ethnicity can only be determined after the behavior is identified as a violation by officers.

REFERENCES


Lamberth, J. (2001, October 22). Testimony Before the House Judiciary Committee’s Subcommittee on Crime and Corrections, Harrisburg, PA.


**Robin Shepard Engel, Ph.D., is an associate professor of criminal justice at the University of Cincinnati. Her current research involves theoretical and empirical explorations of police supervision, patrol officers’ behavior, police response toward problem citizens, and criminal justice theory. She is currently engaged in the collection and analysis of police-citizen contact data during all traffic stops for the Pennsylvania State Police. Recent articles have appeared in Criminology, Justice Quarterly, and Journal of Criminal Justice.**

**Jennifer M. Calnon, M.A., is a doctoral candidate in crime, law, and justice at The Pennsylvania State University. Her research interests include patrol officers’ decision making, the role of race and gender in the criminal justice system, and intimate partner violence. She recently finished her tenure as the project manager for the Project on Police Citizen Contacts with the Pennsylvania State Police. She is currently working on her dissertation research, examining drivers’ speeding behavior in Pennsylvania.**