Understanding Disparities in Policing Outcomes: A Guide to Data Sources, Methods, and Interpretation

March 2026

Prepared by:

Jennifer Cherkauskas, Ph.D., The Ohio State University

Robin S. Engel, Ph.D., The Ohio State University

Nicholas Corsaro, Ph.D., University of Cincinnati

THE OHIO STATE UNIVERSITY

TABLE OF CONTENTS

Executive Summary

Public concern about racial and ethnic disparities in policing outcomes has intensified, driving increased demands for transparency, accountability, and data-driven assessment of police actions. In response, law enforcement agencies are increasingly publishing enforcement data, issuing analytical reports, and partnering with researchers to examine traffic stops, arrests, searches, and use of force. While these efforts are essential, interpreting disparity analyses requires careful attention to data quality, measurement choices, and methodological limitations. This guide is designed to help readers better understand commonly used data sources, statistical techniques, and the interpretation of internal or external law enforcement agency reports on policing outcomes (e.g., traffic stops, arrests, and use of force).

The guide first addresses measurement and units of analysis, emphasizing that how outcomes such as stops, arrests, and use of force are defined and counted fundamentally shapes analytical conclusions. Variations across agencies in reporting requirements, data fields, and units of analysis (e.g., stop-level, subject-level, incident-level, or officer-level) can lead to inconsistent or misleading comparisons if not clearly documented and understood.

Next, the guide reviews key data sources used to examine policing outcomes. Official police administrative data—such as calls for service, stop data, arrests, and use of force reports—form the backbone of most disparity analyses and offer important strengths. Specifically, they systematically and quantitatively document police activity and allow for consistent internal comparisons over time. At the same time, these data are collected for administrative purposes, largely reflect only the officer's perspective, and often lack key contextual detail and temporal sequencing needed to fully explain observed disparities. The guide highlights how additional data sources, such as report narratives, body-worn camera footage, surveys, interviews, and policy reviews, can provide critical context but should be viewed as complementary rather than definitive evidence of disparities.

The core of the guide synthesizes common analytical approaches used to quantitatively assess disparities, including descriptive statistics, bivariate analyses, benchmark analyses, veil-of-darkness tests, interrupted time series analyses, multivariate regression models, predicted probabilities, and outcome tests. For each method, the guide explains what the analysis does, how results should be interpreted, and—critically—what conclusions can and cannot be drawn. A key takeaway is that findings can vary substantially depending on analytical choices, particularly the selection of benchmark populations, and that statistically significant results do not necessarily indicate substantively meaningful differences.

Police officer decision-making is complex and shaped by legal, situational, organizational, and contextual factors that are often not fully captured in administrative data. Therefore, a central premise of the guide is that it is prudent to employ a holistic, multi-method approach to understanding possible disparities in policing outcomes. Each data source, method, and statistical technique offers distinct strengths and limitations. When used as part of a holistic assessment, they collectively provide an opportunity to assess the totality of the evidence toward understanding disparities in policing outcomes more than any single approach.

This is advisable because it is beyond the capacity of any statistical technique to attribute racial differences in policing outcomes to individual officers' or organizational racial bias or discrimination. When used and interpreted appropriately, disparity analyses provide important contextual information and can support transparency, inform supervision and training, identify areas for improvement or further examination, and guide evidence-based organizational change—while avoiding overstatement of what the data can reveal about bias or intent.

1. Introduction

In recent years, public concern about racial and ethnic disparities in policing outcomes has intensified, accompanied by growing demands for transparency, accountability, and data-driven assessment of police actions. This is not a new phenomenon, however, as decades of research have examined the role of individuals' race and ethnicity in criminal justice outcomes.

This literature consistently underscores the absence of a single definitive research method or statistical analysis that is recognized as the best way to determine whether racial and ethnic disparities exist in policing outcomes (Engel & Swartz, 2014; Mears et al., 2016; Sampson & Lauritsen, 1997). Rather, given the complexity of factors that can influence differences across racial/ethnic groups, scholars increasingly emphasize the importance of using multiple methods and data sources to examine disparities to produce a more holistic picture of officer decision-making (Engel & Cherkauskas, 2025; Ratcliffe & Hyland, 2025).

As communities demand more accountability from their police departments, the collection, sharing, and interpretation of enforcement data is a core theme that emerges. We expect police to make decisions in the field based on departmental policy, training, and the law as applied to the specific situations they encounter. We further expect police to be held accountable for their actions through appropriate oversight, supervision, and disciplinary procedures. The systematic collection and appropriate analysis of police data is necessary to identify patterns, trends, and opportunities for improvement.

There is a strong desire from the public to use data to identify racial/ethnic bias in police actions and to hold agencies (and individual officers) accountable when racial profiling, bias[1], and discrimination are detected. The problem, however, is that much of the aggregate data collected on police actions – and the research methodologies and statistical analyses available – cannot meet this public expectation. While official police data can provide a greater understanding of the frequency, context, and circumstances surrounding the use of force (specifically force used against people of color), the limits of aggregate data analyses and quantitative statistics do not allow for the determination of police bias. Police data can identify disparities—that is, differences in outcomes across groups. This can illuminate patterns and trends in several ways (e.g., over time, in comparison with crime and arrest trends, in comparison with other organizational units or geographic areas). However, there is insufficient information to fully determine why these patterns and trends exist. Police decision-making is

[1] For the purposes of this guide, bias is defined as prejudice in favor of or against one thing, person, or group compared with another, usually in a way considered to be unfair.

complex and occurs within dynamic legal, situational, and organizational contexts. Often, this is not what community members, policymakers, and even police executives want to hear, but the fact remains that there are considerable limitations to what researchers can ascertain and conclude about officers' reasoning and motives for using force based on quantitative data.

As calls for police transparency have grown, agencies are increasingly (1) publishing administrative data related to crime, traffic stops, arrests, and use of force on open data portals, (2) summarizing their enforcement data in routine monthly, quarterly, and annual reports, and (3) partnering with outside research teams to conduct more in-depth assessments of data, policy, training, and operations (Bodah & Gilbert, 2022; Brown et al., 2022; Caplan et al., 2015; Chanin & Espinosa, 2015; Morrow, 2021). The primary purpose of annual reports is to describe police activity at an aggregate level by summarizing the frequency and distribution of police outcomes (e.g., traffic stops, arrests, use of force), document patterns and trends, assess changes over time, and explore the potential impact of individuals' race/ethnicity. When used appropriately, these analyses can support transparency, inform internal accountability processes, and help identify areas where policy, training, or supervision may warrant further attention. However, the scope, structure, and analytical approaches of these reports vary widely, as do the conclusions drawn from them. As a result, stakeholders—including community members, policymakers, police executives, and researchers—are often confronted with complex findings that can be difficult to interpret and compare across jurisdictions.

GUIDE OVERVIEW

The following guide is designed to help readers better understand commonly used data sources, statistical techniques, and the interpretation of internal or external law enforcement agency reports on policing outcomes (e.g., traffic stops, arrests, and use of force). The guide first describes issues related to measurement and units of analysis. We then summarize common data sources, the rationale for using each, and their strengths and limitations. Next, the guide explains common methods for analyzing police enforcement data (with a particular focus on understanding racial/ethnic disparities), how to interpret these analyses, the strengths and limitations of each approach, and the conclusions that can and cannot be drawn from them.

Finally, this guide synthesizes findings from peer-reviewed academic articles, independent agency reports, and other best-practice guides (see Appendix A for a select list of resources). We also rely on our expertise and experiences from working directly with numerous law enforcement agencies for over two decades on data collection, analysis, and interpretation related to traffic stops, searches, arrests, and use of force. Our purpose is to provide a better understanding of the complex research methodologies and advanced statistical techniques

often used to guide readers on what these reports can – and cannot – tell us about police enforcement activity, thereby promoting more informed and responsible interpretation of these findings.

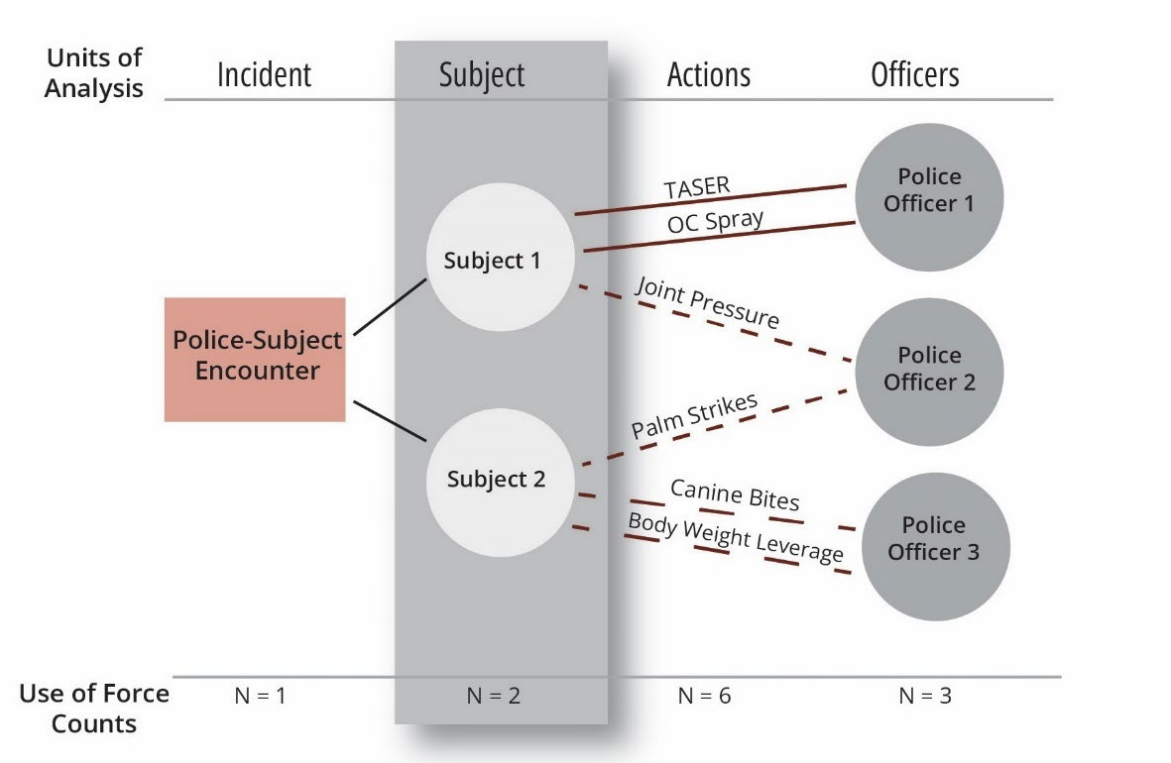
2. Measurement and Units of Analysis

As aptly noted by the Police Executive Research Forum, "agencies cannot manage what they do not measure" (2021, p.5). Measurement begins by clearly defining what you want to study (Maxfield & Babbie, 2014). Once the concept is defined, the next step is to "operationalize" or decide how to represent that concept numerically (Singleton et al., 2005). This process of assigning values is typically how measurement is defined. These measurement decisions matter because they determine which analyses can be conducted. Furthermore, when studies measure the same concept in different ways, they can produce inconsistent findings. As Geller and colleagues argued, "it is important to understand how measurement and other analytic choices affect our understanding of equity in police practices" (2021, p.1083).

To understand patterns and trends in enforcement activity, police agencies must first collect sufficient data. Most agencies are required by policy to systematically report traffic stops, searches, arrests, and use of force. However, reporting requirements or what "counts" as reportable vary dramatically across agencies. For example, some agencies report all traffic stops regardless of disposition, while others only document stops where a citation was issued (NCSL, 2025; Pryor et al., 2020). Most analyses of stop data treat the stop as the unit of analysis, but other possible units of analysis include the driver, the passenger, the officer, specific outcomes (e.g., citations, arrests, searches), or aggregated geographic, organizational, or temporal units.

How force is "counted" also varies across agencies and studies (Hollis, 2018). Some use of force actions are uniformly reported across most police agencies (e.g., the use of physical force, chemical spray, impact weapons, conductive energy devices, and firearms), while there is considerably more variation across agencies in reporting other actions, including verbal commands, threats to use force, handcuffing, escort holds, displaying or pointing of less-lethal weapons, and displaying or pointing of firearms (Fridell, 2017; Geller et al., 2021; Klahm et al., 2014; Pate & Fridell, 1995; Terrill et al., 2018; Willits & Makin, 2018; Wolf et al., 2009). Furthermore, the unit of analysis at which force is measured matters, as illustrated in Figure 1. Typically, the most common units of analysis are the subject and incident. However, other research questions may be most appropriate at other units of analysis. For example, an examination of the effectiveness of different types of force (e.g., physical force, TASER, OC spray, etc.) requires a shift to the force action unit of analysis (Brown et al., 2022).

Figure 1. Hypothetical Units of Analysis for Police Use of Force



Reprinted from Engel et al., 2025

3. Data Sources

Analyses of police enforcement activity rely on multiple data sources that vary in purpose, structure, and analytic utility. This section describes the primary types of official police administrative data commonly used in quantitative research, as well as additional official and non-official data sources that can complement these analyses.

OFFICIAL POLICE ADMINISTRATIVE DATA

Many studies of police enforcement activity rely solely on quantitative analyses of official administrative data. These data sources include police calls for service, pedestrian and traffic stops and searches, reported crime, arrests, and use of force. Table 1 provides an overview of each type of police administrative data, including a brief description and its uses. Using multiple data sources, when appropriate, can provide a more holistic picture of police enforcement activity and outcomes.

Table 1. Official Data Sources for Examining Policing Outcomes

Data Source	Description
Calls for Service (CFS)	• Data documenting dispatched and self-initiated police activity • Dispatched CFS are generated by community members for police assistance (typically through 911 or a non-emergency number) that call takers assign for police response • Self-initiated CFS are a measure of officers' proactive work, where the "call" is generated based on their observations • Can include information on racial/ethnic characteristics of crime suspects recorded through Computer Aided Dispatch (CAD) data • Used to examine temporal and geographic demand for police assistance, and officer-initiated activity; serve as benchmark comparison for other outcomes (Engel et al., 2012; Klinger & Bridges, 1997; Maxfield & Babbie, 2014; Sherman et al., 1989; Smith et al., 2022)
Traffic Stops, Pedestrian Stops, & Searches	• Data collected by police officers documenting their interactions with members of the public during officer-initiated traffic or pedestrian stops • Whether it is collected for all stops or only stops with specific outcomes varies by agency • Data fields vary dramatically by agency, but typically include information on racial/ethnic and other demographic characteristics of the individual stopped, stop characteristics, reason for stop, and outcome of stop (including whether a search was conducted, the reason for search, and whether any contraband was seized) • Used as the "numerator" in benchmark analyses and the primary data source for multivariate analyses predicting stop outcomes and outcome test for seizures (NCSL, 2025; Pryor et al., 2020)

Reported Criminal Offenses	• Crimes reported to the police, which are compiled and reported to the FBI
	• Two primary reporting systems: UCR and NIBRS
	  ○ Uniform Crime Reports (UCR): focuses on 8 part I offenses (murder, rape, robbery, assault, motor vehicle theft, larceny, and burglary) and follows a hierarchy rule where only the highest offense is reported
	  ○ National Incident Based Reporting System (NIBRS): focuses on the incident characteristics for 52 Group A offenses; no hierarchy rule
	• Limitations of both are that they do not capture unreported crimes
	• Typically can be aggregated to incident and suspect levels
	• Used to examine trends in criminal incidents, provide context to overall enforcement activities, and facilitate benchmark comparisons to known criminal suspects.
	
	(Addington, 2019; Asher, 2024; Maxfield & Babbie, 2014)

Data Source	Description
Arrests	• Arrest reports completed by officers when an individual is taken into police custody with the intent of charging them with a crime • Typically documents information about the incident, suspect, criminal offenses/charges, and arresting officer • Typically can be aggregated to the incident and individual arrestee levels • Used to examine arrest trends, examine racial/ethnic disparities, create benchmark comparisons for individuals who experience use of force, and identify factors that predict use of force against arrestees (multivariate regression analyses). (Fridell, 2004; Walker & Katz, 2025)
Use of Force	• Use of force incident reports completed by involved officers and/or their supervisors; typically documented in an electronic reporting system but may be paper reports • Use of force reports document the specific incident and context surrounding the use of force, including the force used and characteristics of the individual against whom force was used. • Typically can be aggregated to the incident, subject, force action, and officer levels • Data most often analyzed at the subject level • Used to examine trends in officers' use of force and facilitate benchmark comparisons using force data as the numerator (Fridell, 2017; Geller et al., 2021; PERF, 2021)
Personnel Data	• Information about police officers, including their demographic characteristics, training, assignments, and other employment records • Typically linked to other official data by badge or employee identification number to measure differences across groups of officers (e.g., officers' rank, age, race, gender, specialized assignment or training)

These types of official data are most useful when captured in an electronic format that allows officers to report information systematically into databases that can be readily quantified, analyzed, searched, and linked to one another (Maxfield & Babbie, 2014; PERF, 2021). The strengths and limitations of using official police data collected by officers are summarized below (Maxfield & Babbie, 2014).

Strengths

- Measures frequency and severity of actual police behavior
- Data are often readily available and quantifiable
- Provides an official "count" that can be systematically compared (within the agency) across officers, organizational units, and time

- Documents information related to involved individuals, situational factors, and legal considerations

Limitations

- Reflect only the officer perspective of the event (except for CFS)
- Collected for administrative and statutory purposes, not research purposes
    - Often, information about factors that can explain officer decision making are either not collected or not readily available in a quantitative format (e.g., only reported in report narratives), including whether a weapon was present, mental health or substance use impairment, whether de-escalation techniques were attempted, etc.
- Typically, do not allow for temporal ordering or sequencing of officer and community member actions during encounter

ADDITIONAL DATA SOURCES

In addition to official administrative data sources, other (non-quantitative) official and non-official data sources can be used to supplement quantitative analyses of administrative reports on enforcement activity.

NON-QUANTITATIVE OFFICIAL DATA SOURCES

First, important contextual information is often included in report narratives (written by officers or supervisors). These written supplements, however, are not readily usable for analytic purposes because they lack uniformity in preparation and are difficult to extract from reports in an efficient manner. While report narratives provide significant additional context that may be later systematically coded and included in quantitative databases, this process is labor-intensive and is rarely conducted within police agencies.

Recent research suggests that advances in natural language processing and machine learning can extract structured, analytical information from unstructured police report narratives while substantially reducing the time and labor required for manual coding (Martin et al., 2023; Relins et al., 2025). However, comparative evidence indicates that NLP tools depend on narrative quality, are better suited to capturing high-level themes, and lack the depth and contextual richness of traditional qualitative analysis (Lukmanjaya et al., 2026; Somers et al., 2025). As a result, these methods are best viewed as tools to supplement—rather than replace—traditional data sources and human review.

Second, the widespread adoption of body-worn camera (BWC) technology provides new opportunities for more detailed assessments of police encounters by systematically reviewing

and coding BWC footage to capture factors not routinely documented in official stop, arrest, or use of force reports. This is a time and resource-intensive methodology that is not routinely conducted within police agencies but is increasingly used by researchers (Terrill & Zimmerman, 2022; Terrill et al., 2023; Worden et al. 2025).

BWC footage allows researchers to directly observe and document officer and subject actions, communication, and the sequence of events leading up to enforcement outcomes (Terrill & Zimmerman; Terrill et al., 2023; Worden et al., 2025). These measures can help contextualize observed disparities in quantitative analyses of administrative data by going beyond whether a specific final outcome occurred to offer potential explanations for why it occurred and whether encounters involving different racial groups differ in observable ways beyond what is recorded in official reports. Video-based observation also avoids influencing officer behavior during the encounter and allows multiple researchers to review the same footage, improving consistency and reliability. Coders can also pause, rewind, and replay footage to increase accuracy in their data collection without having to rely on field notes or memory.

At the same time, BWC footage has clear limitations that influence how findings should be understood (Terrill & Zimmerman, Terrill et al., 2023; Worden et al., 2025). Cameras record events from a limited angle and might miss important actions, environmental cues, or behaviors that happen outside the camera's view or are masked by noise or poor lighting. Some features—such as tone, demeanor, or emotional state—are hard for coders to interpret consistently from video, and footage cannot fully reflect officers' perceptions, intentions, or decision-making. Coding BWC footage is also highly time and resource intensive and requires extensive training to ensure reliability. Therefore, BWC-based coding is best used to complement analyses of quantitative enforcement data by adding context and insights into dynamic interactions, rather than as a standalone tool for making definitive judgments about disparities or bias.

NON-OFFICIAL DATA SOURCES

Often, analyses of official police data can be supplemented by assessing non-official data sources or by using qualitative methods to provide additional contextual information. These additional data sources include 1) surveys, 2) focus groups, 3) interviews, 4) reviews of training and relevant policies, and 5) systematic observation of behavior during ride-alongs. These data sources are briefly summarized in Table 2.

These additional data sources, however, do not provide direct evidence of the presence or absence of racial disparities in policing outcomes and should not be interpreted as doing so. The findings based on these additional data sources provide a broader context for understanding patterns and trends in police enforcement activity, can identify practices to

sustain, and inform where there are opportunities for improvement. Because the primary focus of this guide is on interpreting analyses of racial/ethnic disparities using official police data, the remaining sections address only such data.

Table 2. Non-Official Data Sources for Contextualizing Analyses of Policing Outcomes

Data Source	Description

Survey Data	• Measures attitudes, perceptions, and self-reported behaviors (typically officers and supervisors, but also used for community members) either at a single time (cross-sectional) or over time (waved/longitudinal)
	• Administration methods vary (phone, mail, paper, or online surveys)
	• Strengths
	  ○ Responses from large sample of respondents increase generalizability
	  ○ Community surveys give voice to those with lived experiences
	  ○ Officer surveys gather information from the research subjects themselves
	• Limitations
	  ○ Usually do not measure behavioral outcomes
	  ○ Accuracy of perceptions can't be independently confirmed
	  ○ Can be costly to obtain representative sample of community members
	  ○ External validity/generalizability limited when response rates are low or use non-representative sample
	
	(Maxfield & Babbie, 2014, Nix et al., 2019; Rosenbaum et al., 2017; Weisberg, 2008)

Focus Groups	• A moderated group interview that follows a pre-established protocol designed to elicit discussion on specific topics
	• Emphasizes interaction between participants as part of knowledge-producing process
	• Strengths
	  ○ Can provide rich contextual information for explaining behavior/outcomes of interest beyond the capability of aggregate, quantitative data
	  ○ Behavior is being explained by the research subjects themselves
	  ○ Useful for exploratory research
	• Limitations
	  ○ Reliability: Accuracy of perceptions or self-reported behavior can't be independently confirmed
	  ○ External validity: Participants may not be representative of larger population
	  ○ Groupthink: Participants may not disclose information in front of others
	
	(Krueger & Casey, 2015; Maxfield & Babbie, 2014; Morgan, 1988, 1996)

Interviews	• Structured or semi-structured, but both are designed for the interviewer to gain knowledge by eliciting descriptive information from the interviewee
	• Structured: limited to predetermined, standardized questions and answer sets
	  ○ Useful when the researcher is interested in comparing responses across interviews
	  ○ Limited in their ability to explore topics in-depth or spontaneously

Data Source	Description
	• Semi-structured: standardized questions are the starting point ○ Strength of semi-structured is that it allows the interviewer to ask spontaneous or follow-up questions to explore a topic in more depth (Maxfield & Babbie, 2014)

Review of Training & Policies	• Provide a critical understanding of the content shaping how officers are expected to conduct enforcement
	• Important to ensure that recommendations based on findings are actionable and realistic within organizational context
	• Strengths
	  ○ Can aid in measurement issues associated with quantitative analyses
	  ○ Can compare to national models/best practices if available
	  ○ Does not require direct interaction with human subjects
	• Limitations
	  ○ Describes what officers are supposed to do, which may not match behavior in practice
	  ○ Training reviews document the content, duration, and requirements, but not how well the material was understood, retained, or applied
	  ○ Policies and training often change and current versions may not match the period of study
	(Terrill et al., 2018)

Systematic Social Observation	• Structured observational research methodology for documenting behavior in natural settings using standardized protocols
	• Used to code interactions between the police and the public, including the actions of all involved, situational factors, etc.
	• Strengths
	  ○ Comprehensive situational awareness
	  ○ Provides context and insight into discretionary decisions through debriefing
	  ○ Replicable and quantifiable
	• Limitations
	  ○ Reactivity – officers may behave differently in presence of observer
	  ○ Labor-intensive, expensive, and time-consuming, which can limit sample size
	  ○ Relies on single observer so reliability of recall may be limited
	(Mastrofski et al., 2009; McCluskey et al., 2023; Reiss, 1971)

4. Analyzing Police Enforcement Data

OVERVIEW

Understanding the patterns and trends in officer decision-making during public interactions is crucial for both police executives and the public. To estimate the impact of subjects'

race/ethnicity on the likelihood of different police outcomes, researchers and agency analytical staff can utilize a variety of research methods and statistical techniques. Table 3 summarizes common approaches, which are described more fully in the remainder of this section. Each method or statistical technique offers distinct strengths and limitations, and collectively provides a more comprehensive examination of aggregate-level patterns and trends of law enforcement activity.[2] These analytic approaches vary in complexity, and it is widely recognized that most police agencies lack the internal capacity to execute more sophisticated analyses. As a result, the PERF guide and other experts recommend that agencies establish partnerships with researchers who have both the methodological training and experience working with law enforcement agencies to provide more in-depth analyses of official police data (Crawford, 2019; Engel & Henderson, 2013; Engel & Whalen, 2010; Rojek et al., 2012).

It is critical to note from the outset that officer decision-making is complex, and that while statistical analyses of administrative data described below can identify patterns, trends, and possible explanations for racial or ethnic disparities in enforcement outcomes, **they cannot determine whether those differences are attributable to organizational or individual officers' racial bias or discrimination. Drawing this type of conclusion is beyond the capacity of any statistical technique.** Rather, these analyses provide important contextual information and highlight disparities that warrant further examination using additional data, methods, and forms of inquiry, including supervisory oversight and accountability mechanisms.

The remainder of this section provides the following information for each of the methods summarized in Table 3.

- **Description**: what the analysis is and its purpose
- **Methods**: how it is conducted
- **Interpretation**: what the results mean
- **Strengths & Limitations**: the advantages and disadvantages of this approach
- **Overall conclusions**: what this analysis can and cannot tell us

[2] While a variety of statistical techniques can address research questions about patterns and trends, we limit our discussion in this guide to the analyses routinely used to examine the influence of individuals' race/ethnicity on policing outcomes. It is not, however, meant to be an all-inclusive list of potential methodologies and statistical techniques.

Table 3. Overview of Statistical Techniques to Analyze Police Enforcement Data

Research Method / Statistical Technique	Description
Descriptive Statistics	• Summarize quantitative data using counts and percentages; does not account for variations in trends
Bivariate Analyses	• Evaluate the relationship between two variables; does not consider any other factors that might influence that relationship
Benchmark Analyses	• Examine differences in the representation of racial/ethnic groups in enforcement data (e.g., stops, use of force) compared to their representation in a reference or "benchmark" population that should accurately estimate the population *at risk* of that enforcement action • Findings vary dramatically based on the selected benchmark
Veil of Darkness (traffic stops only)	• Alternative to benchmark analyses • Assess the relative difference in the likelihood of drivers being stopped in daylight vs. darkness across racial/ethnic groups by using changes in natural daylight during inter-twilight period
Interrupted Time Series	• Consider how patterns and trends in police enforcement actions fluctuate over time; specifically examine whether they were significantly impacted or disrupted by seminal events, policy changes, or other interventions
Multivariate Logistic Regression Models	• Estimate independent effect of each predictor on an enforcement outcome while holding other factors constant, allowing for estimation of effect of individuals' race/ethnicity on whether enforcement outcomes occur while other factors are statistically controlled • Estimates odds ratio (chances in favor of outcome ranging from 0 to infinity, with 1 representing equal chance). Does not account for impact of unmeasured factors • Interpretation: 1.0 to 1.49 = substantively small; 1.50 to 2.49 = moderate effect; and 2.5 or higher = large effect
Predicted Probabilities	• Measures likelihood of outcome happening, ranging from zero (impossible) to one (certain) • Predicted probabilities reflect chances of an outcome occurring for average person while all other factors are held at their average
Outcome Test (searches only)	• Statistical comparison of contraband seizure rates across racial/ethnic groups • Only appropriate for discretionary searches (not mandatory or consent searches) • Only measures disparities; cannot measure discrimination / bias

DESCRIPTIVE STATISTICS

DESCRIPTION

To begin to understand the patterns and trends in police enforcement actions, the first step is to describe the available data. Descriptive analyses summarize and present outcome counts.

METHODS

Typical descriptive analyses report the mean (average), minimum and maximum values, and standard deviation for a specific measure or variable.

INTERPRETATION

Descriptive analyses should be interpreted as a factual summary of what is observed in the data, but not as evidence of relationships or explanations.

STRENGTHS & LIMITATIONS

Strengths

- Provide overview of key characteristics and patterns within data (e.g., its central tendencies, variability, and distribution) (Witte & Witte, 2015).

Limitations

- Do not provide any explanation for trends

OVERALL CONCLUSIONS

No conclusions about disparities or their possible causes can be drawn from the basic analyses. These analyses are typically used as a precursor to more complex statistical techniques and to illuminate appropriate methodological approaches (Witte & Witte, 2015).

BIVARIATE ANALYSES

DESCRIPTION

Bivariate analyses examine the association between two variables (e.g., race and use of force, race and reason for stop). They offer an initial assessment of the general trends and potential correlations between the predictor and outcome variables before primary analytical techniques are employed.

METHODS

The number of categories being compared and the level of measurement of the variable (e.g., groups, counts, frequencies) determine the appropriate bivariate test. Bivariate analyses are perhaps most commonly based on the Chi-square test, which assesses whether the association between two variables differs significantly from expected values (Sharpe, 2015; Warne, 2020). For example, a chi-square statistic can be used to examine whether subjects' race and police use of force are related by comparing the actual number use of force subjects observed for each racial/ethnic group to the numbers that would be expected if race/ethnicity had no influence on the likelihood of force being used; the greater the difference between observed and expected counts, the stronger the evidence of an association between race/ethnicity and use of force. Other bivariate analyses that may be used to explore potential disparities are *t*-tests and ANOVA statistics. These provide mean-difference estimates between two groups (or, in the case of ANOVA, three or more groups), where the statistical significance test compares the mean differences between the groups to the 'random noise' in those same outcomes that fluctuates across the groups being compared.

INTERPRETATION

Bivariate analyses focus solely on single associations between two variables; there are no additional 'controls' that can be input into bivariate statistical analyses. Chi-square interpretations focus on differences in likelihood across categories that are not random. A statistically significant chi-square statistic indicates a significant association between categorical variables or a significant difference from expected frequencies. For example, a potential Chi-Square association (and subsequent interpretation) would be that Black, White, and Hispanic drivers have different frequencies of 'equipment violations' during traffic stops. T-tests and ANOVA interpretations focus solely on differences in means (averages) across groups. For example, the mean number of arrests for males compared to females (two-group T-Tests) or the mean number of arrests from a group that compares Black, White, and Hispanic drivers sampled in a study (3 groups, or ANOVA statistics).

Statistical significance reflects the confidence level – or how likely it is – that observed differences are real and not a result of random chance or sampling error. This is represented by a *p*-value, with a common standard of 95% confidence. This indicates confidence that the findings are 5% or less likely to result from chance or sampling error (Betensky, 2019). Analyses that result in statistically significant findings do not necessarily equate to practically or substantively significant findings, as small but consistent findings can produce statistically significant results. Further discussion of statistical significance is provided in the multivariate regression section below.

STRENGTHS & LIMITATIONS

Strengths

- Offers initial insight into the relationships between variables

Limitations

- Does not consider any other factors that might influence that relationship (e.g., seriousness of the offense, time, location, etc.), which may also vary by race/ethnicity

OVERALL CONCLUSIONS

No conclusions about disparities or their possible causes can be drawn. Nevertheless, an initial observation of significant bivariate disparities across racial and ethnic groups may warrant further examination of these differences.

BENCHMARK ANALYSES

DESCRIPTION

For descriptive information about the characteristics of individuals who are stopped or experience use of force to be meaningful, it must be compared with data that approximate the "expected" rates of these outcomes for the same groups in the absence of bias. Benchmark analysis is a statistical method used to examine and assess potential disparities in outcomes across racial/ethnic groups by comparing rates to a reference point (or benchmark). The benchmark population should, as accurately as possible, estimate the population **at risk** of various policing outcomes, which is unknown.

METHODS

To examine racial disparities in policing outcomes, one calculates *disparity ratios*, an easily interpretable technique for comparing groups that experienced a policing outcome (e.g., stop, force) to those at risk of the same outcome, relative to the non-Hispanic White population. The calculation of the disparity ratio is a two-step process. The initial step involves calculating a **Disproportionality Index (DI)** by dividing a racial group's *actual* or observed representation in a policing outcome by that group's *expected* representation in the comparison benchmark (e.g., the suspect population). This calculation measures **within-group** differences.

Disproportionality Index (DI) =	Proportion of racial/ethnic groups' *observed* policing outcome
	Proportion of racial/ethnic groups' *expected* policing outcome

Second, the **Disparity Ratio (DR)** can be calculated to measure **between-group** differences by dividing the DI of the minority group by the DI of the majority group.

Disparity Ratio (DR) =	Minority Group's Disproportionality Index
	Majority Group's Disproportionality Index

INTERPRETATION

Disproportionality indices greater than 1.0 indicate that the group experienced police enforcement actions more often than expected given its representation in the benchmark. In contrast, a value less than 1.0 indicates they experienced enforcement actions less often than expected relative to the same benchmark. The larger the DI, the greater the disproportion between the group's actual and expected representation in the selected benchmark.

The disparity ratio is interpreted as the likelihood of an outcome for a person within a specific minority racial/ethnic group relative *to the majority group*. A disparity ratio greater than 1.0 suggests that Black or Hispanic individuals were more likely than their White counterparts to experience police enforcement actions based on the benchmark used, whereas a disparity ratio less than 1.0 indicates the opposite. For example, if the disparity ratio is 2.0, this indicates that the group of interest (minority group) is roughly *two times more likely* to have force used against them in comparison to the majority group (White, non-Hispanic).

STRENGTHS & LIMITATIONS

Strengths

- Some benchmark data sources are readily available
- Easy to calculate and intuitive to the public
- Can calculate within-group and between-group differences
- Examination of multiple benchmarks can be used to examine volatility or consistency across different comparison groups

Limitations

- No benchmark perfectly captures individuals' risk of police contact and enforcement activity (whether a stop, search, arrest, or force), and there is large variation across benchmarks in their ability to do so. A benchmark with higher validity will yield DIs and DRs with greater validity.
- DIs and DRs can be unstable, particularly when the benchmark comprises a small minority population. Specifically, when the denominator is smaller, minor changes in the numerator more heavily influence the disproportionality index.

- There is no agreed-upon value that unequivocally provides a threshold for a determination of disparity or a specific threshold for how much disparity is "too much" (i.e., a point at which disparity can be attributed to racial bias. Differences between observed and expected policing outcomes may be caused by unmeasured factors (Fridell, 2004; Geller et al., 2021).

Table 4 summarizes common benchmarks used to examine disparities in policing outcomes, including a brief description, their most applicable outcomes, and the strengths and limitations of each. As described above, not all benchmarks are equally valid for representing individuals' risk of the frequency and nature of police contact and enforcement activity. Risk is related to differences in the traffic or other law-violating behavior; where and when they live, work, and drive; police priorities at specific times and locations, and any of the legal and situational factors that affect individuals' likelihood of experiencing police enforcement activity (Alpert et al., 2004; Cesario et al., 2019; Fridell, 2004; Nix et al., 2017; Smith et al., 2019; Tregle et al., 2019).

Although widely used and intuitive, Census data benchmarks have repeatedly been shown to be flawed as a measure for representing the population at risk of experiencing police action (Alpert et al., 2004; Fridell, 2004; Geller et al., 2021; Smith et al., 2019, 2022). First, not all people who live in a city or neighborhood face the same "risk" of police enforcement. For example, the risk of being arrested or having force used is influenced by many factors – including involvement in criminal activity – which may not be evenly distributed across the residential population. Second, Census data do not account for the presence or behavior of non-residents. Third, Census data do not measure the types of characteristics shown by research to put individuals at risk of experiencing policing outcomes, including legally relevant behaviors (Engel et al., 2000; Garner et al., 2002; Morgan et al., 2020). Thus, disproportionality indices and disparity ratios based on residential population data must be interpreted with caution.

Non-census-derived benchmarks better approximate the risk of contact with police that could result in enforcement action. For traffic stops, this includes data sources such as citations, accidents, and observations of roadway usage and violating behavior that more accurately estimate driving frequency, quality, location, and legality. For use of force, scholars generally agree that comparison groups based on arrestees[3] and criminal suspects are stronger approximations of the at-risk population than comparisons that rely on residential population (Cesario et al., 2019; Fryer, 2019; Geller et al., 2021; Smith et al., 2021; Tregle et al., 2019).

[3] Most individuals who are subject to use of force are arrested (Davis et al., 2018; Garner et al., 2018; Hickman et al., 2008).

Table 4. Common Benchmarks for Policing Outcomes

Benchmark	Description	Most Applicable Outcomes	Strengths (+) and Limitations (–)
Residential population	• U.S. Census data on demographic characteristics	• Traffic & ped stops • Use of force	+ Readily available – Does not measure factors known to influence risk of policing outcomes – Does not account for non-residents
Traffic citations[4]	• Tickets issued by police during traffic enforcement	• Traffic stops	+ Readily available + Estimate driving behavior – May underestimate disparities if bias in who is cited
Traffic accidents[5]	• Not-at-fault estimates driving pop • At fault estimates traffic law violators	• Traffic stops	+ Readily available + Estimate driving frequency, quality, law-violating behavior – Race not always included – Difficult to have large enough sample
Observations of roadway usage and/or violating behavior[6]	• Stationary or mobile observation of traffic by trained observers	• Traffic stops	+ Can collect data on actual roadway usage AND law-violating behavior – Expensive, time-intensive – Often limited to speeding violations – Based on limited sample of locations – Rely on observers' perceptions, sometimes at a distance
Calls for Service[7]	• Requests for police assistance through 911 that prompt dispatch & response	• Traffic & ped stops	+ Independent of police + Estimates exposure to police – Perceptual errors of suspect characteristics by callers
Arrests	• Reports completed by police during arrest incidents	• Use of force	+ Most UOF subjects are arrested – Not all force results in arrest – May underestimate disparities if bias in who is arrested[8]
Criminal suspects	• Reports completed by police based on public reports	• Traffic & ped stops • Use of force	+ Independent of police – May overrepresent more serious crimes more likely to be reported[9]

[4] See Smith et al., 2021
[5] See Alpert et al., 2004; Lovrich et al., 2007; Smith et al., 2021; Withrow & Williams, 2015; Wolfe et al., 2021
[6] See Alpert et al., 2007; Engel et al., 2003, 2005; Lange et al., 2005; Tillyer & Engel, 2012; Zingraff et al., 2000
[7] See Ratcliffe & Hyland, 2025
[8] See Cesario et al., 2019; Geller et al., 2021; Knox et al., 2020a, 2020b; Knox & Mummolo, 2020
[9] See Klinger & Bridges, 1997

OVERALL CONCLUSIONS

All benchmarks have limitations and vary in the extent to which they accurately estimate the population of similarly situated individuals "at risk" of police enforcement actions (Engel & Calnon, 2004a; PERF, 2021; Ratcliffe & Hyland, 2025; Tillyer et al., 2010). Using the residential population as a comparison benchmark does not account for the likelihood or risk of police enforcement activity and, therefore, is one of the weakest benchmark comparisons. Nevertheless, benchmark comparisons based on Census data can be useful context for two narrow purposes. First, these analyses provide a baseline of how different racial/ethnic groups experience enforcement actions. Second, comparing disparity ratios across a variety of benchmarks helps determine the validity of the analytical technique for representing the population at risk of police enforcement actions.

Studies over the last twenty years have consistently demonstrated that the use of different benchmark populations can result in dramatically different findings (Brown et al, 2022; Engel et al., 2003, 2005; Cesario et al., 2019; Geller et al., 2021; Smith et al. 2019, 2022; Tregle et al., 2019). Much of this research shows that benchmark comparisons based on population statistics nearly always show racial/ethnic disparities in policing outcomes, while benchmarks that better approximate risk show reduced or no racial/ethnic disparities. Therefore, it is critical to know and understand the strengths and limitations of the benchmark population being used. Recently, criminologist Ian Adams posted a free interactive tool called the Disparity Benchmark Simulator that illustrates how assessments of racial disparity depend heavily upon the choice of comparison group (i.e., the benchmark). In sum, based on the limitations and volatility of benchmark analyses, they can only be used to identify the presence of disparities in outcomes, not the presence of bias.

VEIL OF DARKNESS (TRAFFIC STOPS ONLY)

DESCRIPTION

In response to the limitations of benchmark comparisons, Grogger & Ridgeway (2006) developed the Veil of Darkness (VOD) technique. A VOD analysis assesses whether the likelihood of stops of Black or Hispanic drivers significantly differs between daylight and darkness, when driver race is presumed to be more or less visible.

METHODS

The VOD statistical method uses multivariate logistic regression to estimate whether a stop involves a Black or Hispanic driver. By analyzing a subset of traffic stops that occur during the "inter-twilight period," the VOD approach leverages natural variations in daylight throughout the year. This allows the assessment of relative differences in the ratio of minority to non-

minority stops in daylight versus darkness. The VOD approach does not assert that identifying drivers' characteristics at night is impossible or that it is always feasible during the day; instead, it posits that recognizing driver characteristics is generally more challenging in the dark (Grogger & Ridgeway, 2006; Knode et al., 2024).

INTERPRETATION

The VOD technique is based on multivariate logistic regression models that produce odds ratios for each independent variable (see further description below). With *daylight* as the variable of interest and controls for seasonality, day of the week, time of day, the results can be interpreted as the likelihood of a Black or Hispanic driver being stopped in the daylight versus the darkness.

STRENGTHS & LIMITATIONS

Strengths

- Reduces need to rely solely on benchmark comparisons in assessing disparities in initial stop decisions
- Uses natural experiment based on seasonal variations in daylight hours to determine whether traffic stops of drivers of color differ when driver race is more or less visible

Limitations

- Focuses on a smaller subset of traffic stops, limiting its generalizability to stops made at other times
- Darkness is an imperfect proxy for officers' ability to identify driver race (e.g., window tint, speed, exterior lighting, etc.)
- Assumes racial composition of drivers using and violating traffic laws does not differ between daylight and darkness

OVERALL CONCLUSIONS

The VOD approach can determine whether disparities in traffic stops exist in a way that aligns with how visible driver race is assumed to be at the time of the stop. However, it cannot explain whether any observed disparities are due to differences in officer deployment or enforcement priorities, driving behavior, or racial bias because officers' intent and ability to identify drivers are not directly measured.

INTERRUPTED TIME SERIES

DESCRIPTION

Interrupted time series (ITS) modeling is a statistical technique that provides a robust quasi-experimental counterfactual framework, particularly when treatment-to-control matching is not possible or when targeted outcome measures are unavailable in non-treatment settings (Cook & Campbell, 1979). ITS assesses whether there are statistically significant differences in the average frequency of an outcome (i.e., use of force or other event counts) across repeated measures of the outcome that correspond with period-specific events or interventions (i.e., a before/after design with a designated, distinct, and unique pre/post period).

METHODS

Different interrupted time series methods are appropriate for different types of outcomes, with model choice and data requirements varying based on how frequently events occur and how they are distributed over time. When using interrupted time series analyses, the choice of statistical method depends largely on how the outcome data behave over time. If outcomes are relatively common and follow a pattern where average values are fairly stable (that is, the mean, median, and most frequent values are similar), researchers often use Autoregressive Integrated Moving Average (ARIMA) models (McCleary et al., 1980). These methods are well suited for continuous or frequently occurring outcomes but generally require a large number of observations—typically at least 50 time periods—to produce reliable results (Box & Tiao, 1975).

When outcomes are less frequent and measured as counts (such as rare events per month), Poisson regression is often a more appropriate and stable approach (Kuhn et al., 1994). The number of observations needed for Poisson-based time series analyses depends on how large a change is expected following an intervention or policy shift. Larger anticipated changes require fewer observations to detect statistically, while smaller changes require more data.[10] In practice, most interrupted time series analyses rely on somewhere between 50 and 150 time periods, with roughly 60 to 90 observations often sufficient to assess meaningful changes in outcomes.

[10] For example, detecting a large change in outcomes (around 40 percent) may require fewer than 100 time periods, whereas detecting more modest changes (around 30 percent) may require closer to 150 observations.

INTERPRETATION

The important point about interpreting time series findings (regardless of whether the results are derived from ARIMA or ML estimation) is to understand that the estimated change is simply the mean difference in the outcome of interest between the pre- and post-intervention (or event) periods, while simultaneously controlling for time. In many ways, the time series are analogous to a *t*-test of mean difference, controlling for time-varying factors, to show whether the specific timing of the modeled event/intervention was associated with a significant shift in the outcome of interest.[11]

STRENGTHS & LIMITATIONS

Strengths

- Useful for examining outcomes measured as monthly counts
- Controls for time-varying factors
- Particularly useful for examining the impact of seminal events, policy changes, etc.

Limitations

- Require a long enough period following the date of the "interruption" to determine if there are statistically significant changes to the overall time trend. Using a short post-time period can lead to unstable statistical estimates.
- If multiple events are in close time proximity, cannot fully disentangle their independent impact on the frequency of use of force
- Not particularly well suited to determining whether there is a change in ratios or proportions of events for different racial/ethnic groups

OVERALL CONCLUSIONS

When examining racial and ethnic disparities in outcomes, a natural question that frequently emerges is – did this unique and distinct event date (e.g., high-profile critical incident), or intervention date (e.g., policy change date) correspond with differences in race-related outcomes. Time series analyses can offer insight into whether a consistently measured outcome of interest (e.g., use-of-force counts, civilian or officer injuries, total arrests, etc.) across racial and ethnic groups shifted at distinct points in time, provided specific assumptions

---

[11] The one technical difference for interpretation is contingent upon the modeling technique used. ARIMA estimates are simply *raw mean changes* (e.g., a significant estimate of –5.65 would be 5.65 fewer events in the designated period modeled – such as months – in the post-intervention period relative to the pre-intervention period). For ML estimation, the scale is based on the natural logarithm scale – so estimates are interpreted as *percentage changes* in the outcome in the post-period relative to the pre-period.

are met.[12] Time series analyses provide evidence that transitional events can correspond with (and possibly lead to) changes in outcomes of interest.

MULTIVARIATE REGRESSION ANALYSES

Given the widely recognized limitations of benchmark analyses, most examinations of police use of force and other outcomes (e.g., stops, citations, arrests) also employ multivariate regression analysis. Unlike benchmark analyses that are necessary because the pool of eligible drivers to be stopped is uncertain, one significant benefit of analyzing post-stop enforcement outcomes is that the population of those stopped is known. Similarly, the likelihood of force being used against a known population (e.g., individuals arrested) can be estimated using multivariate regression modeling.

Consequently, more robust statistical and methodological approaches can be utilized to investigate any racial or ethnic disparities in enforcement actions that happen after the initial stop is made. These analyses seek to answer the question: What factors influence the likelihood of receiving a warning, citation, arrest, or search? This is particularly useful because multiple factors can affect officers' decision-making during interactions with the public. For instance, characteristics of the driver, the vehicle, the stop itself, the reasons for the stop, other legal considerations, and the officer's characteristics have all been shown in previous studies to impact post-stop enforcement results (Engel & Calnon, 2004b; Schafer et al., 2006; Tillyer et al., 2019; Tillyer & Engel, 2013).

DESCRIPTION

Multivariate regression models estimate the independent effect of each predictor variable in the model on the outcome (e.g., citation, arrest, force, injury, etc.), while controlling for the predictive power of all other variables in the model. For example, it can estimate the likelihood of use of force among arrested individuals and isolate the impact of key variables of interest (e.g., race/ethnicity).

[12] There are different ways to structure time series analyses, and some designs are more appropriate than others. For example, studies may include a roughly equal number of observations before and after an intervention, or a longer pre-intervention period followed by a shorter post-intervention period (such as 6–12 observations). However, it is not methodologically appropriate to rely on very few observations before an intervention and many observations afterward, as this violates key assumptions of time series analysis and can lead to misleading results.

METHODS

Multivariate regression modeling creates a mathematical equation that simultaneously considers the influence of multiple variables on an outcome, thus providing a more nuanced understanding and insights into the interplay of factors contributing to racial disparities.[13] The occurrence or non-occurrence of specific enforcement outcomes during a police encounter with a member of the public typically yields a binary result, indicating that the outcome of interest is dichotomous. In these circumstances, logistic regression is the appropriate statistical modeling technique, defined as (0 = does not occur, 1 = does occur) (Hanushek & Jackson, 1977; Liao, 1994; Meyers et al., 2016).

INTERPRETATION

When interpreting multivariate logistic regression models, there are several components to consider. First, these models reveal the relative strength of relationships between the independent variables and the dependent variable through two related metrics for each independent variable: (1) the coefficient, indicating the predicted log-odds, and (2) the odds ratio. The coefficient provides an additive measure of a specific variable's influence. A negative coefficient signifies a negative relationship, suggesting that the variable's impact makes the enforcement outcome less likely. Conversely, a positive coefficient (no sign) implies that the variable increases the likelihood of the enforcement outcome.

In logistic regression models, the estimated effects of the variables are typically expressed as odds ratios, which indicate the strength and direction of the relationship between each factor and the outcome on a standardized scale.[14] An odds ratio greater than one indicates the variable is associated with higher odds of the event occurring, while an odds ratio less than one suggests association with lower odds of that occurrence. Odds ratios indicate how the likelihood of the enforcement outcome changes due to a specific variable. It is crucial to assess the influence of a variable, illustrated by the size of the odds ratio, which reflects the strength of its relationship with the dependent variable. The standard guidance regarding the size of odds ratios suggests that odds ratios less than 1.5 are substantively small, 1.5 to 2.5 are medium, and 2.6 or greater are substantively large (Chen et al., 2010).

---

[13] This description and summary of multivariate logistic regression modeling is derived from various sources (see Hanushek & Jackson, 1977; Liao, 1994; Long, 1997; Meyers et al., 2016; Witte & Witte, 2015).

[14] Technically, this odds ratio represents a type of log-odds; however, interpreting this value can be non-intuitive. For this reason, it is common to exponentiate the coefficient for clearer interpretation in terms of odds (Liao, 1994). The odds ratio reflects this transformation by converting the coefficient into the multiplicative odds of the outcome variable relative to the predictor variable, assuming all other factors remain constant.

Second, significant findings imply statistical significance, indicating a confidence level that the observed differences are not due to random chance or sampling error. While differences might be observed across coefficients, they may not be statistically significant. This means we cannot be confident that the difference is not attributable to random chance. Each variable in the model has a defined significance threshold reflected by a p-value. As noted above, a traditional confidence level of 95% is generally used, meaning the result is 5% or less likely to be due to random chance or sampling error (Betensky, 2019). However, in large samples, significance testing can be more sensitive to very small or artifactual relationships between variables, thus detecting statistically significant differences that lack substantive significance or practical meaning (Allison, 1999; Benjamin & Berger, 2019; Greenland et al., 2016; Wasserstein & Lazar, 2016). Thus, when assessing the impact of specific factors on post-stop enforcement outcomes, it is appropriate to prioritize the size of the regression coefficients and the odds ratios (which reflect the strength of the relationship) over mere statistical significance.

Third, while multivariate statistical modeling is a more comprehensive analytical strategy than bivariate analysis, a significant limitation is that it can only statistically control for measured variables. This limitation is known as "model specification error," which refers to the error in a statistical model due to the inability to account for all factors influencing the outcome (Hanushek & Jackson, 1977; Jung et al., 2018; Marvell & Moody, 1996). In official police data collection systems, it is not possible for officers to systematically document information on every relevant factor that might explain officer decision-making and enforcement outcomes. Unmeasured or unincluded variables can potentially bias estimates and results.

The Nagelkerke R-square statistic may also be presented for each model. This measure, relevant to binary logistic regression, offers a general view of model goodness-of-fit. In the social sciences, a common guideline indicates that a model with an R-square less than .10 is considered poorly fitting, one between 0.10 and 0.20 is viewed as a weak-to-solid fitting model, and those above 0.20 are regarded as robust fitting models (Muijs, 2012). The fitting of the model provides a loose interpretation of how strong the model is for predicting the outcome of interest (i.e., use of force).

STRENGTHS & LIMITATIONS

Strengths

- Isolate the individual impact of multiple factors simultaneously
- Identify the significant and most substantive factors to explain outcomes that can prioritize administrators' exploration of findings
- Can use predicted probabilities to more precisely estimate the impact of independent variables in the multivariate model (see below)

Limitations

- Omitted variable problem – can only control for predictors that are measured and available in the data, so the model has not properly eliminated other potential explanations for the outcome
- May not be able to establish time order or sequencing with official data

OVERALL CONCLUSIONS

Greater confidence can be placed in the interpretation of findings produced by multivariate models as compared to bivariate analyses. However, due to the inability to collect data on all factors relevant to officer enforcement decision-making, these findings should be interpreted with this inherent limitation in mind.

PREDICTED PROBABILITIES

A related approach to multivariate regression analysis is the use of predicted probability analyses, which more precisely estimate how variables in multivariate regression models impact a specific outcome (e.g., impact of driver or subject race/ethnicity on enforcement outcomes). As noted above, the "odds" indicate whether an outcome is more or less likely to occur for one group compared to a reference group (with values ranging from zero to infinity, where "1" indicates equal chances. In contrast, predicted probabilities measure the expected chances that an outcome will happen given a specific set of conditions (e.g., Black driver or White suspect), ranging from zero (impossible) to one (certain), while controlling for the remaining factors in the model. Predicted probabilities offer a more accurate risk assessment than the general outcome percentage, given the models' accuracy and predictive capabilities.

METHODS

The use of predicted probabilities derived from logistic regression estimation follows a standard three-step process (Liao, 1994). First, the linear component of a regression model is converted to log-odds (to estimate the log-odds of an event occurring or not occurring based on the estimated regression). Second, the log-odds are converted to predicted odds (to compare the relative impact of each of the regression coefficients in the model with one another (but still on the logarithm scale). Third, the log-odds are converted back to the original scale by converting the predicted log-odds into a probability that ranges between 0 (the closer a predicted probability is to 0, the less likely it is to happen) and 1 (the closer to 1 a predicted probability is, the more likely it is to happen).

INTERPRETATION

The predicted probabilities for stop outcomes reflect the chance of an event occurring for an average person/stop/encounter, while considering all variables in the models. For example, calculating the probabilities for White, Black, and Hispanic drivers based on various stop-related situational and legal factors allows for a comparison of estimates among different racial and ethnic groups regarding their *probability* of warning, citation, arrest, or search, assuming all else is equal (i.e., all other measures in the models are set to their mean) values.

Essentially, predicted probabilities answer whether a specific trait of interest (e.g., a person's race/ethnicity) changes the probability that an event (e.g., an arrest) will occur, net of all other control variables.

STRENGTHS & LIMITATIONS

Strengths

- Isolate the individual impact of multiple factors simultaneously
- More precisely estimates the chances of specific outcomes than the odds ratios produced by the multivariate regression model

Limitations

- Omitted variable problem – can only control for predictors that are measured and available in the data, so the model has not properly eliminated other potential explanations for the outcome

OVERALL CONCLUSIONS

Finally, as noted above for multivariate regression analyses, we can have more confidence in the results of predicted probabilities analyses than bivariate or benchmark analyses. However, model misspecification remains a limitation to consider when interpreting these findings.

OUTCOME TEST (TRAFFIC OR PEDESTRIAN STOPS ONLY)

DESCRIPTION

Identifying contraband during searches of individuals and vehicles is a key outcome when exploring possible racial or ethnic disparities. Commonly known as search “success rates” or “hit rates” (i.e., the percentage of searches that yield contraband), some researchers apply the “outcome test” to identify these disparities by analyzing differences in search success rates (Ayres, 2001; Chohlas-Wood et al., 2022; Knowles et al., 2001). The application of the outcome test to police searches is based on the premise that if officers profile drivers or pedestrians due

to racial bias, they will persist in searching Blacks and Hispanics even when the likelihood of finding contraband is lower compared to searches of Whites, but if people are searched solely on the basis of legitimate legal factors and suspicions unrelated to race, similar percentages of searches resulting in seizures should be expected across racial groups (Anwar & Fang, 2006; Ayres, 2001; Knowles et al., 2001). These scholars suggest that, in the absence of bias, a state of equilibrium will eventually be achieved, whereby police searches among racial groups are proportional to their actual possession of contraband. The reliance on the principle of equilibrium eliminates the need for incorporating multiple variables (i.e., a multivariate model).

METHODS

The outcome test relies on comparisons of seizure rates across racial and ethnic groups (a crosstabulation) and is calculated as the percentage of searches in which officers seize contraband (e.g., drugs, illegal weapons) for each group relative to the total number of searches conducted for each group (Fridell, 2004; Ramirez et al., 2000).

INTERPRETATION

The outcome test interprets differences in contraband seizure rates across racial and ethnic groups as evidence of disparities in search outcomes. For example, if 30% of searches of White drivers resulted in a contraband seizure, but only 20% of searches of Black drivers did, this would be interpreted as searches of Black drivers being less likely to result in the discovery of contraband than searches of White drivers.

STRENGTHS & LIMITATIONS

Strengths

- Does not require an external benchmark population because the population at risk is known (i.e., all who were searched)
- Offers an alternative method to multivariate regression analyses for assessing post-stop searches (i.e., robust to the model misspecification problem)
- Intuitive interpretation
- Can be complemented by the robust outcome test,[15] which strengthens inference by requiring consistent disparities in both search rates and seizure rates

[15] Recently, Gaebler and Goel (2025) introduced the "robust outcome test," which they argue improves upon the standard outcome test by examining whether there are disparities in *both* an officer's decision to search *and* the seizure rate. If a group is searched more often but those searches are less successful, it suggests that officers may

Limitations[16]

- Assumes officers possess complete discretion over conducting searches[17]
- Assumes officers do not consider motorists' or pedestrians' behaviors when deciding on searches
- Assumes all officers' search decisions are uniform
- Assumes the sole purpose of a search is to uncover contraband, ignoring officer safety
- Assumes the different racial groups have the same risk distributions (e.g., likelihood of carrying contraband)

OVERALL CONCLUSIONS

The outcome test is intended to assess patterns at an aggregate level and does not measure individual officer intent or bias. Differences in seizure rates may also reflect factors not captured in the analysis. Findings from the outcome test should therefore be interpreted cautiously and in combination with other analyses and contextual information. No definitive conclusions about racial bias should be drawn based on racial or ethnic disparities in seizure rates identified using this method.

5. CONCLUSION

As shown throughout this guide, examining and understanding disparities in policing results is inherently complex and requires a careful, methodologically sound approach. Analyzing policing outcomes has important implications for police departments and communities, making responsible data interpretation especially important. It is critical to fully acknowledge the methodological and analytical strengths and limitations of available data and methods.

Statistical analyses of aggregate police administrative data are valuable for identifying patterns and trends in enforcement activity, including pinpointing outliers by geography, organizational unit, or encounter type (PERF, 2021). While these patterns may identify disparities that warrant further attention, they can also reflect a range of unmeasured factors

---

be using a lower standard of evidence to search that group, which points to possible discrimination. If the two measures (search rate and seizure rate) do not clearly point in the same direction, the test results are inconclusive.

[16] See Engel, 2008; Engel & Tillyer, 2008; and Simoiu et al., 2017 for more detailed discussions of the assumptions and limitations related to the outcome test.

[17] Based on this criterion, the outcome test is suitable solely for examining traffic stops that lead to a probable cause or reasonable suspicion search. Mandatory searches should be excluded because officers are required to conduct them under specific conditions. Consent searches are more complex. Although officers initially decide from whom to request consent to search, it is ultimately the motorists who decide whether the consent searches take place (Dias et al., 2024; Engel, 2008; Fridell, 2004). Motorists have the right to refuse search requests, and if an officer lacks probable cause, they must respect the denial.

and should not be interpreted as definitive evidence of individual bias or racially biased policing. Examining enforcement data across time and organizational contexts can highlight evidence-based opportunities to strengthen supervision, refine policy, and enhance training.

Aggregate data analyses can also serve an important baseline function, allowing agencies to assess change over time and evaluate the impact of policy, training, or operational updates or interventions. When conducted routinely and transparently, these regular assessments of operations through quantitative analyses can support internal accountability, inform continuous improvement efforts, and provide agencies with a framework for communicating progress to the public.

However, even the most comprehensive data collection systems and rigorous statistical analyses have inherent limitations. No single data source or statistical method can fully capture the complexity of police-public interactions, but a combination of varied approaches maximizes the strengths of each (Brent & Kraska, 2010; Trahan & Stewart, 2013; Worrall, 2000). Taken individually, singular approaches to examining enforcement activity will provide only a partial picture of the relationships between race/ethnicity and police actions and insufficient information upon which to improve agency operations.

A holistic approach involves using multiple data sources, measures, methods, and analytical techniques. It allows police agencies and researchers to address interrelated questions by understanding patterns and trends over time (descriptive analyses, interrupted time series analyses), identifying and quantifying observed disparities (benchmark analyses), and identifying possible contextual factors that contribute to police enforcement actions (multivariate regression analysis). Supplementing analyses of administrative data with the collection and analysis of other quantitative data (e.g., surveys) and qualitative methods (e.g., focus groups, interviews) provides critical insight into how and why patterns emerge. A holistic approach provides a more comprehensive understanding of enforcement activity, officers' decision-making, and observed disparities; it also supports the development of more targeted and effective organizational responses.

In sum, while statistical analyses of administrative data can reveal the presence of racial/ethnic disparities in outcomes and begin to explain why at least some of them exist, they cannot 1) determine the legality of individual encounters with police, or 2) determine whether they result from discriminatory intent or practice at the individual or organizational level. These limitations are well-recognized in the scholarly literature, but not always clearly communicated to or understood by the public (Engel & Calnon, 2004a; Fridell, 2004; Pryor et al., 2020; Tillyer et al., 2010). It is important that readers avoid emphasizing isolated metrics or selectively highlighting results that align with a particular narrative. Instead, assessing the totality of the evidence is advisable. When findings converge across

different data sources and methods, confidence in conclusions is strengthened. However, a measured interpretation of all analytical results is still prudent. When findings diverge, ambiguity should be acknowledged rather than obscured. Further attention and in-depth examination are needed to clarify. Ultimately, this balanced and transparent approach supports the development of actionable, agency-specific recommendations to reduce disparities in enforcement actions.

In conclusion, the main takeaway from this guide is that analyses of racial disparities in policing enforcement outcomes must be approached with both analytical rigor and appropriate caution. Responsible interpretation and reliance on a holistic methodological approach are essential for translating evidence into improvements in policing practice.

Appendix A: List of Resource Guides & Scholarly Research

TRAFFIC STOPS, PEDESTRIAN STOPS, AND SEARCHES

RESOURCE GUIDES

Fridell, L. (2004). *By the Numbers: A Guide for Analyzing Race Data from Vehicle Stops.* Police Executive Research Forum, Washington, DC. By the Numbers

Pryor, M., Goff, P.A., Heydari, F., & Friedman, B. (2020). *Collecting, Analyzing, and Responding to Stop Data: A Guidebook for Law Enforcement Agencies, Government, and Communities*. https://policingequity.org/images/pdfs-doc/COPS-Guidebook_Final_Release_Version_2-compressed.pdf.

SCHOLARLY ARTICLES

Alpert, G. P., Dunham, R. G., & Smith, M. R. (2007). Investigating racial profiling by the Miami Dade Police Department: A multimethod approach. *Criminology & Public Policy*, *6*(1), 25-55.

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

Dias, M., Epp, D. A., Roman, M., & Walker, H. L. (2024). Consent searches: Evaluating the usefulness of a common and highly discretionary police practice. *Journal of Empirical Legal Studies*, *21*(1), 35-91.

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

Engel, R.S. & Calnon, J. (2004a). Comparing Benchmark Methodologies for Police-Citizen Contacts: Traffic Stop Data Collection for the Pennsylvania State Police. *Police Quarterly*, *7*(1), 97–125.

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

Engel, R.S. & Tillyer, R. (2008). Searching for equilibrium: The tenuous nature of the outcome test. *Justice Quarterly, 25*(1), 54-71.

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

Knode, J. L., Wolfe, S. E., & Carter, T. M. (2024). Pulling back the veil of darkness: A proposed road map to disentangle racial disparities in traffic stops, a research note. *Criminology*, *62*(2), 364–375. https://doi.org/10.1111/1745-9125.12366

Ratcliffe, J. H., & Hyland, S. S. (2025). Police stops and naïve denominators. *Crime Science, 14*(1), 10.

Smith, M. R., Tillyer, R., Smith, M., & Lloyd, C. D. (2022). Assessing police stops of pedestrians: Toward a new generation of benchmarks. *Urban Affairs Review*, *58*(4), 1152-1181.

Tillyer, R., Engel, R. S., & Calnon Cherkauskas, J. (2010). Best practices in vehicle stop data collection and analysis. *Policing*, *33*(1), 69.

ARRESTS

SCHOLARLY ARTICLES

Engel, R. S., Smith, M. R., & Cullen, F. T. (2012). Race, place, and drug enforcement: Reconsidering the impact of citizen complaints and crime rates on drug arrests. *Criminology & Public Policy*, *11*, 603.

Gase, L. N., Glenn, B. A., Gomez, L. M., Kuo, T., Inkelas, M., & Ponce, N. A. (2016). Understanding racial and ethnic disparities in arrest: The role of individual, home, school, and community characteristics. *Race and Social Problems*, *8*(4), 296-312.

Huff, J. (2021). Understanding police decisions to arrest: The impact of situational, officer, and neighborhood characteristics on police discretion. *Journal of Criminal Justice*, *75*, 101829.

Kochel, T. R., Wilson, D. B., & Mastrofski, S. D. (2011). Effect of suspect race on officers' arrest decisions. *Criminology*, *49*(2), 473-512.

Neil, R., & MacDonald, J. M. (2023). Where racial and ethnic disparities in policing come from: The spatial concentration of arrests across six cities. *Criminology & Public Policy*, *22*(1), 7-34.

USE OF FORCE

RESOURCE GUIDES

Police Executive Research Forum. (2021). *What police chiefs and sheriffs need to know about collecting and analyzing use-of-force data.* Washington, DC: Police Executive Research Forum. Available: https://www.policeforum.org/assets/CollectingAnalyzingUOFData.pdf

SCHOLARLY ARTICLES

Fridell, L. (2017). Explaining the disparity in results across studies assessing racial disparity in police use of force: A research note. *American Journal of Criminal Justice, 42*, 502-513.

Geller, A., Goff, P. A., Lloyd, T., Haviland, A., Obermark, D., & Glaser, J. (2021). Measuring racial disparities in police use of force: methods matter. *Journal of Quantitative Criminology, 37*(4), 1083-1113.

Tregle, B., Nix, J., & Alpert, G.P. (2019). Disparity does not mean bias: Making sense of observed racial disparities in fatal officer-involved shootings with multiple benchmarks. *Journal of Crime and Justice, 42*, 18-31.

Willits, D. W., & Makin, D. A. (2018). Show me what happened: Analyzing use of force through analysis of body-worn camera footage. *Journal of Research in Crime and Delinquency, 55*(1), 51-77.

References

Addington, L.A. (2019). NIBRS as the New Normal: What Fully Incident-Based Crime Data Mean for Researchers. In: Krohn, M., Hendrix, N., Penly Hall, G., Lizotte, A. (eds) Handbook on Crime and Deviance. Handbooks of Sociology and Social Research. Springer, Cham.

Allison, P.D. (1999). *Multiple Regression: A Primer*. Thousand Oaks, CA: Sage Publications.

Alpert, G.P, Dunham, R.G., & Smith, M.R. (2007). Investigating racial profiling by the Miami-Dade Police Department: A multimethod approach. *Criminology & Public Policy, 6(*1), 201-232.

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

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

Asher, J. (2024, July 29). *Another Cautionary Crime Data Tale*. https://jasher.substack.com/p/another-cautionary-crime-data-tale-289

Ayres, I. (2001). *Pervasive Prejudice? Unconventional Evidence of Racial and Gender Discrimination.* Chicago: The University of Chicago Press.

Benjamin, D., & Berger, J. (2019). Three recommendations for improving the use of p-values. *The American Statistician, 73*, 186-191.

Betensky, R. (2019). The p-Value Requires Context, Not a Threshold. *The American Statistician*, 73(1), 115-117. DOI: 10.1080/00031305.2018.1529624

Bodah, D. & Gilbert, D. (2022). *The Police Data Transparency Index*. Vera Institute. https://policetransparency.vera.org/PTI-factsheet.pdf

Box, G. E., & Tiao, G. C. (1975). Intervention analysis with applications to economic and environmental problems. *Journal of the American Statistical Association*, *70*(349), 70-79.

Brent, J.J. & Kraska, P.B. (2010). Moving beyond our methodological default: A case for mixed methods. *Journal of Criminal Justice Education, 21*(4), 412-430.

Brown, R., Engel, R.S., Cherkauskas, J., Corsaro, N., & Kurtz, J. D. (2022). Assessment of Colorado Springs Police Department Use of Force. Report submitted to the Colorado Springs

Police Department, Office of the Chief. https://coloradosprings.gov/sites/default/files/inline-images/cspd_use_of_force_final_transparency_matters_report_april_2022.pdf

Caplan, R., Rosenblat, A., & Boyd, D. (2015). Open data, the criminal justice system, and the police data initiative. *Data and civil rights: A new era of policing and justice. Washington, DC: Data*, 1-13.

Cesario, J., Johnson, D. J., & Terrill, W. (2019). Is There Evidence of Racial Disparity in Police Use of Deadly Force? Analyses of Officer-Involved Fatal Shootings in 2015–2016. *Social psychological and personality science*, *10*(5).

Chanin, J., & Espinosa, S. (2016). Examining the determinants of police department transparency: The view of police executives. *Criminal Justice Policy Review*, *27*(5), 498-519.

Chen, H., Cohen, P., & Chen, S. (2010). How big is a big odds ratio? Interpreting the magnitudes of odds ratios in epidemiological studies. *Communications in Statistics—simulation and Computation, 39*(4), 860-864.

Chohlas-Wood, A., Gerchick, M., Goel, S., Huq, A. Z., Shoemaker, A., Shroff, R., & Yao, K. (2022). Identifying and measuring excessive and discriminatory policing. *U. Chi. L. Rev.*, *89*, 441-475.

Cook, T.D. & Campbell, D.T. (1979). *Quasi-Experimentation: Design and Analysis Issues for Field Settings.* Boston: Houghton Mifflin.

Crawford, A. (2019). Effecting change in policing through police/academic partnerships: The challenges of (and for) co-production. In *Critical Reflections on Evidence-Based Policing* (pp. 175-197). Routledge.

Davis, E., Whyde, A., & Langton, L. (2018). *Contacts between police and the public, 2015*. Washington, DC: U. S. Department of Justice.

Engel, R., & Calnon, J. (2004). Comparing Benchmark Methodologies for Police-Citizen Contacts: Traffic Stop Data Collection for the Pennsylvania State Police. *Police Quarterly*, *7*(1), 97–125.

Engel, R.S., Calnon, J.M., Tillyer, R., Johnson, R., Liu, L., & Wang, X. (2005). *Project on Police-Citizen Contacts: Year 2 Final Report.* Submitted to the Pennsylvania State Police Department, Harrisburg, PA.

Engel, R.S., Calnon, J.M., Liu, Lin, & Johnson, R. (2003). *Project on Police-Citizen Contacts: Year 1 Final Report.* Submitted to the Pennsylvania State Police Department, Harrisburg, PA.

Engel, R.S. & Cherkauskas, J.C. (2025). *2024 Pennsylvania State Police Traffic Stop Study: January 1 – December 31, 2024*. Report submitted to the Commissioner of the Pennsylvania State Police.

Engel, R.S., Cherkauskas, J.C., Corsaro, N.C., Yildirim, M., McManus, H., & Fisher, R. (2023). *Enforcement data analysis for the Aurora, Colorado Police Department*. Report submitted to IntegrAssure, LLC, the City of Aurora, and the Aurora Police Department.

Engel, R. S., Corsaro, N., Motz, R. T., & Cherkauskas, J. (2025). *Evaluation of Integrating Communications, Assessment, and Tactics (ICAT) Training with the Indianapolis Metropolitan Police.* Submitted to the National Institute of Justice, Washington DC, and the Office of the Chief of Police, Indianapolis Metropolitan Police Department, Indianapolis, IN. https://glenn.osu.edu/sites/default/files/2025-08/NIJ_IMPD_ICAT_Evaluation_Final.pdf

Engel, R. S., & Henderson, S. (2013). Beyond Rhetoric: Establishing police—academic partnerships that work. In *The future of policing* (pp. 217-236). Routledge.

Engel, R. S., Sobol, J. J., & Worden, R. E. (2000). Further exploration of the demeanor hypothesis: The interaction effects of suspects' characteristics and demeanor on police behavior. Justice Quarterly, 17(2), 235–258.

Engel, R. S., Smith, M. R., & Cullen, F. T. (2012). Race, place, and drug enforcement: Reconsidering the impact of citizen complaints and crime rates on drug arrests. *Criminology & Public Policy*, *11*, 603.

Engel, R. S., & Swartz, K. (2014). Race, crime, and policing. *The Oxford handbook of ethnicity, crime, and immigration*, 135-165.

Engel, R. S., & Whalen, J. L. (2010). Police–academic partnerships: Ending the dialogue of the deaf, the Cincinnati experience. *Police Practice and Research: An International Journal, 11*(2), 105-116.

Fridell, L. A. (2004). *By the Numbers: A Guide for Analyzing Race Data from Vehicle Stops*. Washington, DC: Police Executive Research Forum.

Fryer Jr., R. G. (2019). An empirical analysis of racial differences in police use of force. *Journal of Political Economy*, *127*(3), 1210-1261.

Gaebler, J. D., & Goel, S. (2025). A simple, statistically robust test of discrimination. *Proceedings of the National Academy of Sciences*, *122*(10), e2416348122.

Garner, J. H., Hickman, M. J., Malega, R. W., Maxwell, C. D. (2018). Progress toward national estimates of police use of force. *PLoS ONE*, *13*(2), e0192932.

Garner, J. H., Maxwell, C. D., & Heraux, C. G. (2002). Characteristics associated with the prevalence and severity of force used by the police. *Justice Quarterly*, *19*(4), 705–746.

Geller, A., Goff, P. A., Lloyd, T., Haviland, A., Obermark, D., & Glaser, J. (2021). Measuring racial disparities in police use of force: methods matter. *Journal of Quantitative Criminology*.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology, 31* (4), 337-350.

Hanushek, E. A., & Jackson, J. E. (1977). *Statistical methods for social scientists*. Orlando, FL: Academic Press.

Hickman, M. J., Piquero, A. R. & Garner, J. H. (2008). Toward a national estimate of police use of nonlethal force. *Criminology & Public Policy, 7*(4), 563-604.

Hollis, M. E. (2018). Measurement issues in police use of force: A state-of-the-art review. *Policing, 41*(6), 844-858.

Jung, J., Corbett-Davies, S., Shroff, R., & Goel, S. (2018). Omitted and included variable bias in tests for disparate impact.

King, G. (1988). Statistical Models for Political Science Event Counts: Bias in Conventional Procedures and Evidence for the Exponential Poisson Regression Model. *American Journal of Political Science, 32*(3), 838–863. https://doi.org/10.2307/2111248

Klahm IV, C. F., Frank, J., & Liederbach, J. (2014). Understanding police use of force: Rethinking the link between conceptualization and measurement. *Policing: An International Journal of Police Strategies & Management*, *37*(3), 558-578.

Klinger, D. A., & Bridges, G. S. (1997). Measurement error in calls-for-service as an indicator of crime. *Criminology*, *35*(4), 705-726).

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

Knox, D., W. Lowe, and J. Mummolo. (2020a). Administrative Records Mask Racially Biased Policing. *American Political Science Review, 114*(3), 619-637.

Knox, D., Lowe, W., & Mummolo, J. (2020b). Can Racial Bias in Policing Be Credibly Estimated Using Data Contaminated by Post-Treatment Selection?. *Available at SSRN 3940802*.

Knox, D., & Mummolo, J. (2020). Toward a general causal framework for the study of racial bias in policing. *Journal of Political Institutions and Political Economy, 1*(3), 341-378.

Krueger, R.A. & Casey, M.A. (2015). *Focus Groups: A Practical Guide for Applied Research*. *5th Edition*. Thousand Oaks, CA: SAGE.

Kuhn, L., Davidson, L. L., & Durkin, M. S. (1994). Use of Poisson regression and time series analysis for detecting changes over time in rates of child injury following a prevention program. *American Journal of Epidemiology*, *140*(10), 943-955.

Lange, J. E., Johnson, M. B., & Voas, R. B. (2005). Testing the racial profiling hypothesis for seemingly disparate traffic stops on the New Jersey Turnpike. *Justice Quarterly, 22*(2), 193-223.

Liao, T.F. (1994). Interpreting Probability Models: Logit, Probit, and Other Generalized Linear Models. Thousand Oaks, CA: Sage.

Long J. S. (1997). Regression models for categorical and limited dependent variables. *Sage Advanced Quantitative Techniques in the Social Sciences Series*. Vol. 7. Thousand Oaks, CA: Sage.

Lovrich, N.P., Gaffney, M.J., Mosher, C.C., & Pratt, T.C. (2007). *Results of the monitoring of WSP traffic stops for biased policing.* Washington State University.

Lukmanjaya, W., Halmich, C., Butler, T., Cook, D., & Karystianis, G. (2026). Computational text analysis on unstructured police data: A scoping review. *Crime Science, 15*, Article 272. https://doi.org/10.1186/s40163-026-00272-2

Martin, A., Fokoue, E., Fu, H., & Tessier, H. (2023). *Learning from inaccessible data: Natural language processing on police use of force reports*. Office of Community Oriented Policing Services, U.S. Department of Justice. https://portal.cops.usdoj.gov/resourcecenter/content.ashx/cops-r1135-pub.pdf

Marvell, T. B., & Moody, C. E. (1996). Specification problems, police levels, and crime rates. *Criminology*, *34*(4), 609-646.

Mastrofski, S. D., Parks, R. B., & McCluskey, J. D. (2009). Systematic social observation in criminology. In *Handbook of quantitative criminology* (pp. 225-247). New York, NY: Springer New York.

Maxfield, M.G. & Babbie, E. (2014). *Research Methods for Criminal Justice and Criminology*. 7th Ed. Cengage Learning.

McCleary, R., Hay, R. A., Meidinger, E. E., & McDowall, D. (1980). *Applied time series analysis for the social sciences* (Vol. 10). Beverly Hills, CA: Sage Publications.

McCluskey, J., Uchida, C. D., Feys, Y., & Solomon, S. E. (2023). *Systematic social observation of the police in the 21st century*. Springer.

Mears, D. P., Cochran, J. C., & Lindsey, A. M. (2016). Offending and racial and ethnic disparities in criminal justice: A conceptual framework for guiding theory and research and informing policy. *Journal of Contemporary Criminal Justice*, *32*(1), 78-103.

Meyers, L. S., Gamst, G., & Guarino, A. J. (2016). *Applied multivariate research: Design and interpretation.* Sage Publications.

Morgan, D.L. (1996). Focus groups. *Annual Review of Sociology, 22*,129-152.

Morgan, D.L. (1988). *Focus group as qualitative research*. Newbury Park, CA: Sage Publications.

Morgan, M. A., Logan, M. W., & Olma, T. M. (2020). Police use of force and suspect behavior: An inmate perspective. *Journal of Criminal Justice*, *67*, 101673.

Morrow, B. (2021, Nov 29). Law Enforcement Agencies Benefit from Transparent Crime Data. *StateTech Magazine* https://statetechmagazine.com/article/2021/11/law-enforcement-agencies-benefit-transparent-crime-data

Muijs, D. (2012). Advanced quantitative data analysis. In *Research Methods Educational Leadership and Management* (Briggs, A., Coleman, M., & Morrison, M., Eds). Sage Publications, pp: 363-380.

National Conference of State Legislatures (2025). *Law Enforcement Stop Data Collection Database*. https://www.ncsl.org/civil-and-criminal-justice/traffic-stop-data

Nix, J., Pickett, J. T., Baek, H., & Alpert, G. P. (2019). Police research, officer surveys, and response rates. *Policing and society*, 29(5), 530-550.

Pate, A. M., & Fridell, L. A. (1995). Toward the uniform reporting of police use of force: Results of a national survey. *Criminal Justice Review, 20*(2), 123-145.

Police Executive Research Forum. (2021). *What Police Chiefs and Sheriffs Need to Know About Collecting and Analyzing Use-of-Force Data*. https://www.policeforum.org/assets/CollectingAnalyzingUOFData.pdf

Pryor, M., Goff, P.A., Heydari, F., & Friedman, B. (2020). *Collecting, Analyzing, and Responding to Stop Data: A Guidebook for Law Enforcement Agencies, Government, and Communities*. https://policingequity.org/images/pdfs-doc/COPS-Guidebook_Final_Release_Version_2-compressed.pdf.

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.

Ratcliffe, J. H., & Hyland, S. S. (2025). Police stops and naïve denominators. *Crime Science, 14*(1), 10.

Reiss, A. J. (1971). Systematic observation of natural social phenomena. *Sociological methodology*, *3*, 3-33.

Relins, S., Birks, D., & Lloyd, C. (2025). Using instruction-tuned large language models to identify indicators of vulnerability in police incident narratives. *Journal of Quantitative Criminology, 41*, 647–684. https://doi.org/10.1007/s10940-025-09611-z

Ridgeway, G., & MacDonald, J. (2010). Methods for assessing racially biased policing. *Race, ethnicity, and policing: New and essential readings*, 180-204.

Rojek, J., Smith, H. P., & Alpert, G. P. (2012). The prevalence and characteristics of police practitioner–researcher partnerships. *Police Quarterly, 15*(3), 241-261.

Rosenbaum, D. P., Maskaly, J., Lawrence, D. S., Escamilla, J. H., Enciso, G., Christoff, T. E., & Posick, C. (2017). The Police-Community Interaction Survey: measuring police performance in new ways. *Policing: an international journal of police strategies & management*, *40*(1), 112-127.

Ross, C. T., Winterhalder, B., & McElreath, R. (2020). Racial Disparities in Police Use of Deadly Force Against Unarmed Individuals Persist After Appropriately Benchmarking Shooting Data on Violent Crime Rates. *Social Psychological and Personality Science*.

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

Schafer, J. A., Carter, D. L., Katz-Bannister, A. J., & Wells, W. M. (2006). Decision making in traffic stop encounters: A multivariate analysis of police behavior. *Police Quarterly*, *9*(2), 184-209.

Sharpe, D. (2015). Chisquare test is statistically significant: Now what? *Practical Assessment, Research & Evaluation, 20*(1), 1–10. https://doi.org/10.7275/tbfa-x148

Sherman, L. W., Gartin, P. R., & Buerger, M. E. (1989). Hot spots of predatory crime: Routine activities and the criminology of place. *Criminology, 27*(1), 27-56.

Simoiu, C., Corbett-Davies, S., & Goel, S. (2017). The problem of infra-marginality in outcome tests for discrimination. *The Annals of Applied Statistics*, 1193-1216.

Singleton, R.A., Jr., Straits, B.C., & Straits, M.M. (2005). *Approaches to Social Research, 4th ed*. New York: Oxford University Press.

Smith, M. R., Tillyer, R., & Engel, R. S. (2022). Race and the use of force by police revisited: Post-Ferguson findings from a large county police agency. *Police Quarterly, 26*(4), 411-440.

Smith, J. M., Tillyer, R., Engel, R. S., & Cherkauskas, J. C. (2019). A Multi-Method Investigation of Officer Decision-Making and Force Used or Avoided in Arrest Situations: Tulsa, Oklahoma Police Department Administrative Data Analysis Report. Cincinnati, OH: IACP/UC Center for Police Research and Policy.

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

Smith, M. R., Tillyer, R., Smith, M., & Lloyd, C. D. (2022). Assessing police stops of pedestrians: Toward a new generation of benchmarks. *Urban Affairs Review*, *58*(4), 1152-1181.

Somers, L. J., Todak, N., Mourtgos, S. M., & Adams, I. T. (2025). Through Thick and Thin: Comparing Traditional Qualitative Analysis and Natural Language Processing Techniques Using Narrative Data from Police Officers. *Justice Quarterly*, 1-21.

Terrill, W., Paoline, E. A., & Ingram, J. R. (2018). Beyond the final report: A research note on the assessing police use of force policy and outcomes project. *Policing, 41*(2), 194-201.

Terrill, W., & Zimmerman, L. (2022). Police use of force escalation and de-escalation: The use of systematic social observation with video footage. *Police Quarterly, 25*(2), 155–177. https://doi. org/10.1177/10986111211049145

Terrill, W., Zimmerman, L., & Somers, L. J. (2023). Applying video-based systematic social observation to police use of force encounters: An assessment of de-escalation and escalation

within the context of proportionality and incrementalism. *Justice Quarterly, 40*(7), 1045–1076. https://doi.org/10.1080/07418825.2023.2222819

Tillyer, R., & Engel, R. S. (2012). Racial differences in speeding patterns: Exploring the differential offending hypothesis. *Journal of Criminal Justice*, *40*(4), 285-295.

Tillyer, R., & Engel, R. S. (2013). The impact of drivers' race, gender, and age during traffic stops: Assessing interaction terms and the social conditioning model. *Crime & Delinquency*, *59*(3), 369-395.

Tillyer, R., Engel, R. S., & Calnon Cherkauskas, J. (2010). Best practices in vehicle stop data collection and analysis. *Policing*, *33*(1), 69.

Tillyer, R., Smith, M., & Lloyd, C. D. (2019). Another piece of the puzzle: The importance of officer characteristics and group processes in understanding post-stop outcomes. *Journal of Research in Crime and Delinquency*, *56*(5), 736-779.

Todak, N., & James, L. (2018). A systematic social observation study of police de-escalation tactics. *Police Quarterly*, *21*(4), 509-543.

Trahan, A., & Stewart, D. M. (2013). Toward a pragmatic framework for mixed-methods research in criminal justice and criminology. *Applied Psychology in Criminal Justice*, *9*(1), 59-74.

Tregle, B., Nix, J., & Alpert, G.P. (2019). Disparity does not mean bias: Making sense of observed racial disparities in fatal officer-involved shootings with multiple benchmarks. *Journal of Crime and Justice, 42*, 18-31.

Walker, S., & Katz, C. M. (2025). *The police in America: An introduction* (10th ed.). McGraw Hill Education.

Warne, R. T. (2020). ChiSquared test. In *Statistics for the social sciences: A general linear model approach* (pp. 410–455). Cambridge University Press.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on p-values: Context, process, and purpose. *The American Statistician, 70*, 129-133.

Weisberg, H. F. (2008). The methodological strengths and weaknesses of survey research. In W. Donsbach, M.W., Traugott, & H.F., Weisberg, H.F. (Eds,), *The SAGE Handbook of Public Opinion Research* (pp. 223-231). Sage Publications.

Withrow, B.L., & Williams, H. (2015). Proposing a benchmark based on vehicle collision data in racial profiling research. *Criminal Justice Review, 40*(4), 449-469.

Witte, R. S., & Witte, J. S. (2015). *Statistics*. John Wiley & Sons.

Wolf, R., Mesloh, C., Henych, M., & Thompson, L. F. (2009). Police use of force and the cumulative force factor. *Policing: An International Journal of Police Strategies & Management, 32*(4), 739-757.

Wolfe, S. E., Carter, T., & Knode, J. (2021). Michigan State Police Traffic Stop External Benchmarking: A Final Report on Racial and Ethnic Disparities. East Lansing, MI: School of Criminal Justice, Michigan State University. https://justiceresearch.dspacedirect.org/server/api/core/bitstreams/07e3639b-d881-4b67-ac95-8c8e8d5b5ae5/content

Worden, R. E., Holladay, B. P., McLean, S. J., Cochran, H., & Reynolds, D. L. (2025). Systematic social observation of police-citizen encounters: Coding and measurement through body-worn cameras. *Justice Quarterly*, *42*(7), 1410-1443.

Worrall, J.L. (2000). In defense of the "quantoids": More on the reasons for the quantitative emphasis in criminal justice education and research*, Journal of Criminal Justice Education, 11*(2), 353-361.

Zingraff, M., Smith, W., & Tomaskovic-Devey, D. (2000). North Carolina highway traffic and patrol study: Driving while black. *The Criminologist*, *25*(3), 1-4.