EXPLAINING RACE GAPS IN POLICING: NORMATIVE AND EMPIRICAL CHALLENGES

SONJA B. STARR

Explaining Race Gaps in Policing: Normative and Empirical Challenges

Sonja B. Starr*
January 19, 2015

This piece explores the many kinds of quantitative claims that researchers and commentators regularly make about race and policing. Everyone agrees that there are enormous racial gaps in U.S. rates of stops, arrests, searches, and use of force. But there are dramatically conflicting claims as to why. Policing is hard to study, but the problem isn't just the data shortcomings with which the literature has long struggled. It's confusion about what questions we should be asking. Different kinds of numerical comparisons and research designs often imply sharply differing conceptions of what racial equality in policing means. These normative premises often go unstated, such that readers may easily miss these differences. The overarching objective of this Article is to highlight the connection between the normative and the empirical. I identify plausible conceptions of racial equality in policing and assess which empirical methods can best test those conceptions.

The Article gives particular attention to how researchers should address two important research questions. The first is whether criminal conduct differences explain policing disparities. Empirical researchers as well as casual commentators typically purport to address this question either by comparing racial groups' shares of police interactions to their shares of crime, or by comparing two groups' ratio of police interactions to their ratio of crimes. Using examples and mathematical proofs, I show that neither of these comparison types answers the key question whether people with like criminal conduct are being treated the same way. These comparisons generally overcorrect for racial differences in criminal conduct, misleadingly masking the size (or even reversing the apparent direction) of disparities in policing of people with the same *conduct. Second, I examine how researchers should investigate the effects of racial discrimination—a morally important and legally central question, but one that poses serious causal inference challenges. I review several methods in the current literature, which offer useful insights but have substantial limitations, and critique the recently dominant "hit-rate" approach, which relies on faulty normative and empirical premises. Instead, I propose supplementing existing tools with a new approach: the use of "testers."*

* Professor of Law, University of Michigan; Visiting Professor of Law, Harvard University. Thanks to Alicia Davis, Avlana Eisenberg, Mark Fancher, Jim Greiner, Sam Gross, Louis Kaplow, Randy Kennedy, David Moran, J.J. Prescott, Eve Brensike Primus, Jon Sacks, Margo Schlanger, Michael Steinberg, Kim Thomas, and participants in the Michigan Law Faculty Brownbag for helpful comments and conversations. Brian Apel, Grady Bridges, Alex Harris, Avi Kupfer, Linfeng Liu, and Andrew Sand provided excellent research assistance.

INTRODUCTION

As recent events have painfully illustrated, public debates over race and policing are typically catalyzed by flashpoints—individual, terrible cases, often involving police killings of unarmed civilians. But these debates are shaped by competing underlying understandings of the everyday realities of law enforcement. On average, people of color in the United States, especially black men, interact with police far more often than white Americans do. Black men are 2.5 times as likely to be arrested as white men, and local studies show even larger gaps in stops, searches, and use of force, though there are no national numbers.[1] The existence of these gaps is not contested; the reasons are. Do these patterns reflect race-based targeting, or differences in criminal conduct? Or are there other contributing factors—for example, are citizens more likely to report crimes with minority suspects?

Such questions sharply divide public opinion—largely along racial lines[2]—and among public commentators, polar opposite answers are each often presented as essentially indisputable.[3] This dissensus does not merely result from one side or another ignoring or twisting facts. Rather, these questions also have no clear answer in the large empirical literature. The problem goes beyond heterogeneity across locations and police forces, and beyond well-recognized data limitations. It stems in part from normative and conceptual confusions that suffuse the field. Researchers often do not articulate exactly what question they seek to answer and why policymakers should care about it. Sometimes, they pose one question but use an empirical model that effectively answers another.

This Article is an effort at clarity. It is written from an empiricist's perspective, but isn't an empirical paper and doesn't answer the questions posed above. Instead, it addresses threshold questions about research objectives and design: First, what should policing-disparity studies seek to estimate, and why? Second, what empirical strategies can best identify those quantities of interest? I hope this discussion will be useful not just for researchers, but also for policymakers, judges, and citizens who wish to make sense of the bewildering array of statistics on race and policing and to recognize when those statistics are misleading.

There is no single answer to the question of what researchers should estimate; the answer depends on the purpose of the research. But policing-disparity researchers typically seek to inform policy or legal debates in some way, so they

---

[1] *See infra* notes 7-22 and accompanying text (discussing raw disparity statistics).

[2] *E.g.*, Ronald Weitzer & Steven A. Tuch, *Racially Biased Policing: Determinants of Citizen Perceptions*, 83 SOCIAL FORCES 1009, 1017 (2005) (finding blacks are six times likelier than whites to believe racial profiling is a problem).

[3] For example, a recent letter from dozens of civil rights and community leaders called the pattern of young black men subjected to "aggressive police tactics…too obvious to be a coincidence and too frequent to be a mistake…[I]t is time for our country to counter the effects of systemic racial bias." Letter from Maya Rockeymoore et al., to President Obama (Aug. 25, 2014), *available at* http://www.washingtonpost.com/wp-srv/ad/public/static/letter/. By contrast, prominent commentator Heather Mac Donald stated: "It is black crime rates that predict the presence of blacks in the criminal justice system. Not some miscarriage of justice." Meet the Press, Aug. 17, 2014.

should focus on some objective that matters (or *should* matter) to policymakers or lawyers. When such researchers specify their empirical models, they are not just making technical decisions. Rather, their choices imply normative judgments about what racial equality in policing means and what inequalities are worth studying. When researchers do not explicitly explain those judgments, it's easy for legal and policy commentators to misunderstand the studies' implications. Sometimes, researchers themselves seem to share those misunderstandings.

To illustrate these problems in more depth, much of this Article focuses on two types of inequality that hold particular policy or legal interest: (1) disparity in police interactions conditional on criminal conduct (that is, holding criminal conduct constant); and (2) the causal effects of police racial discrimination. These are not the only worthwhile targets of empirical research, but they are important ones.

Extensive research and commentary focuses solely on whether race gaps in policing are explained by crime, ignoring other potential explanatory variables. This literature contains a rich and important debate about how to measure crime, and I offer thoughts on the competing methods. But I focus chiefly on unexamined problems concerning what researchers should do with whatever crime measure they settle on—that is, what does it mean to "account for crime"?

To explore this question, one first must ask why so many scholars and commentators focus exclusively on crime's explanatory role. The answer isn't just its descriptive importance. Rather, the shared assumption seems to be that crime differences could potentially *justify* policing differences in a way that other explanations cannot. The most plausible reasons for this assumption imply a particular equality objective: people with the same criminal conduct should face the same probability of police interactions, regardless of race. This principle is an instantiation of the moral intuition that like cases should be treated alike.

But the specific types of policing-to-crime comparisons that pervade the literature fail to test whether equality in this sense exists. The two common approaches compare a group's *share* of police interactions to its *share* of crimes, or compare the *ratio* of two groups' police interaction rates to the *ratio* of their crime rates. The assumption is that racially equitable policing would produce a racial distribution of police interactions that mirrors the distribution of crime. The basic problem is that this would only be so if everyone the police interact with is guilty of the crime(s) in question. In reality, though, few police interventions are confined to the guilty. Even if the police are quite good (but not perfect) at targeting the guilty, Share/Share and Ratio/Ratio comparisons can be surprisingly misleading.

As I show with examples and mathematical proofs, when there is racial *equality* in policing of those with like conduct, Share/Share or Ratio/Ratio comparisons always misleadingly suggest *disproportionality*—specifically, that the higher-crime group is being "underpoliced" after accounting for crime. And when the higher-crime group is in fact "overpoliced" on average conditional on criminal conduct, the common comparisons mask those disparities' size, or even reverse their apparent direction. The literature is thus rife with comparisons that overstate the extent to which race gaps in police interactions can be explained by criminal conduct. (Even those arguing

that crime does *not* fully explain policing gaps draw such comparisons, generally with the effect of understating their arguments.) These comparisons seem intuitively sensible, but our intuitions are wrong. While "accounting for crime" is a valid objective, the specific ways crime is routinely "accounted for" exaggerate its explanatory value. I offer thoughts on how we can get the right numbers instead.

The second estimand on which I focus, the effect of police racial discrimination, is central to equal protection doctrine, and is also important for policy purposes; the experience of being targeted because of race is a key reason why communities often see heightened police presence as a burden. But empirically identifying racial discrimination is difficult. It requires disentangling other potential causes of disparity—not just criminal conduct, but other potential confounding variables, some of which may be unobservable to researchers.

The literature has taken a variety of approaches to this problem. Economic literature has recently been dominated by the "hit-rate" approach, which posits that irrational police discrimination can be measured by comparing the rates at which police actions produce evidence of crime. Unfortunately, although this approach is mathematically elegant, it does not tell us anything we should care about. It is unrelated to any defensible conception of either equality or efficiency, relies on implausible empirical assumptions, and makes demonstrably false predictions.

More insight can be gained from various more traditional observational methods, although significant omitted-variables and sample-selection concerns mean that cautious interpretation is required; moreover, many studies use inappropriate controls that themselves reflect discretionary police decisions. Some clever studies exploit variation in decision-makers' access to race information, potentially providing stronger causal identification, but these too have substantial limitations. Finally, lab experiments demonstrating prevalent implicit racial bias support stronger causal inferences, but do not tell us how these biases translate into real-world outcomes.

To supplement these tools, I propose a new approach: "auditing" the police using paired testers of different races. Auditing is a staple of research on (and enforcement of laws against) employment, housing, and lending discrimination. In the policing context, the approach has not been tried, likely due to safety, ethical, and legal concerns. I discuss ways to mitigate these concerns through careful research design and cooperation with police departments or other government agencies. Despite its challenges, this approach offers something that other methods generally do not: the promise of strong causal identification in a real-world setting.

Part I describes raw racial gaps in U.S. police interactions and their relationship to punishment disparities, and introduces normative questions surrounding their empirical assessment. Part II examines how and why the relationship between policing gaps and criminal conduct should be estimated. Part III examines the estimation of police racial discrimination, and proposes the auditing method.

I. Stakes and Objectives

In this Part, I discuss what's at stake in efforts to quantify racial disparities in policing. Section A describes the problem these efforts seek to explain: large race

gaps in criminal justice involvement. I also show that without understanding policing disparities, it is difficult to interpret disparities in later criminal process stages and in incarceration. Section B outlines various possible objectives of empirical estimation.

A. Race and Criminal Justice

Let's begin with what we know: people of color, especially black men, are involved in the U.S. criminal justice system at highly disproportionate rates. These "raw" disparities have been the focus of some empirical work,[4] some scholars' policy arguments,[5] and much media coverage.[6] Herein, I use the term "disparity" to refer to raw outcome gaps, not to any particular reason for those gaps, unless specified.

1. Disparities in Police Interactions

In 2011, the arrest rate for all black adults was approximately 10%; the rate for all white adults was approximately 4%.[7] Rates are especially high for black men.[8] African-Americans on average are arrested on more serious charges: compared to whites, nearly four times as often for violent crime, eight times for murder specifically, four times for drug sales or manufacturing, and 4.5 times for weapons offenses.[9] These figures (and the underlying data) don't differentiate by Hispanic ethnicity; if African-Americans were compared to non-Hispanic whites, the gaps would surely be larger.[10] Because black-white disparities are particularly large, I use them as the paradigmatic research target in this Article. Estimating other demographic disparities poses the same basic challenges.

We lack national data on pre-arrest policing decisions, including traffic and pedestrian stops, frisks, and searches.[11] However, local studies have reported large

---

[4] *E.g.*, Michael R. Smith & Matthew Petrocelli, *Racial Profiling? A Multivariate Analysis of Traffic Stop Data*, 4 POLICE Q. 4, 9-12 (2001); Jeff Rojek et al., *The Influence of Driver's Race on Traffic Stops in Missouri*, 7 POLICE Q. 126 (2004).

[5] *E.g.*, I. Bennet Cappers, *Rethinking the Fourth Amendment: Race, Citizenship, and the Equality Principle*, 46 HARV. C.R.-C.L. L. REV. 1 (2011); Kevin R. Johnson, *How Racial Profiling in America Became the Law of the Land*, 98 GEO. L.J. 1005 (2010); Floyd Weatherspoon, *Ending Racial Profiling of African-Americans in the Selective Enforcement of Laws: In Search of Viable Remedies*, 65 U. Pitt. L. Rev. 721 (2004).

[6] *E.g.*, Jess Bidgood, *Boston Police Focus on Blacks in Disproportionate Numbers, Study Shows*, NEW YORK TIMES, Oct. 8, 2014; *see* Greg Ridgeway & John MacDonald, *Methods for Assessing Racially Biased Policing*, in RACE, ETHNICITY, AND POLICING 181 (2010) (Stephen K. Rice & Michael D. White eds.) (describing "compulsion in media reports" to focus on these comparisons).

[7] These rates come from the online arrest rate calculation tool from the Bureau of Justice Statistics, available at http://www.bjs.gov/index.cfm?ty=datool&surl=/arrests/index.cfm#.

[8] *Id.* The BJS data are not broken down by race and sex combined, but in general, men are more than three times as likely to be arrested as women are.

[9] *Id.*

[10] Hispanic incarceration rates are nearly twice those of non-Hispanic whites, Leah Sakata, *Breaking Down Mass Incarceration in the U.S. Census*, Prison Policy Initiative, May 28, 2014, so arrest rates are presumably also higher. In the ethnically undifferentiated data, far more Hispanics are likely described as "white" than as "black." *See* U.S. Census Bureau Working Paper, *America's Churning Races*, 16 tbl. 1.

[11] *E.g.*, Rachel Harmon, *Why Do We (Still) Lack Data on Policing?*, 96 MARQ. L. REV. 1119, 1129-30 (2013). *See also* Samuel R. Gross & Katherine R. Barnes, *Road Work: Racial Profiling and Drug Interdiction on the Highway*, 101 MICH. L. REV. 651, 678-82 (2002) (observing that police can fudge data, such as by "ghosting," reporting stops of white drivers that never occurred).

gaps. For example, "Blacks were subjected to 63% of [police-pedestrian] encounters, even though they made up just 24% of Boston's population."[12] Police data fairly consistently show that black and Hispanic drivers are disproportionately stopped and searched.[13] Data on use of force are limited, but fatalities are well documented. From 1976 to 1998, African-Americans were four times as likely as whites to be killed by police.[14] A new, federally funded initiative is developing a national Justice Database covering stops and use of force.[15]

2. Incarceration Disparities and their Relationship to Policing

Raw disparities in U.S. incarceration rates are even larger. Black men are incarcerated at six times the rate of non-Hispanic white men (and fifty times that of white women); Hispanic men fall midway between. Because the U.S. has an exceptionally high overall incarceration rate,[16] these gaps translate into astonishing total numbers. One in fifteen adult black men is currently behind bars, including one in nine under age 35.[17] The lifetime incarceration hazard for black males is approximately 1 in 3.[18] When probation and parole figures are added, one-third of black men under 35 are *currently* under criminal justice supervision.[19]

Researchers have generally attributed the majority of incarceration disparity to arrest differences.[20] So to understand the reasons for incarceration gaps, we need to understand why there are such sharp racial differences in arrests. Without this understanding, we also may know less than we think we know about disparities in subsequent process stages, such as sentencing. Studies of later process stages use samples consisting only of cases that made it into the criminal justice system and use control variables (such as arrest offense or conviction offense) that are shaped by police officers' earlier decisions. Police discrimination could introduce sample selection bias and could distort the control variables as well.

To illustrate, assume that white and black crime patterns are identical, but that police discriminate against blacks, such that other factors equal, police are more likely to arrest black suspects, or to arrest them on more serious charges (for example, describing an assault as "aggravated"). If so, the black arrestee pool for any

---

[12] ACLU, "Black, Brown, and Targeted: A Report on Boston Police Department Street Encounters from 2007-2010," at 1 (2014), https://www.aclum.org/sites/all/files/images/education/stopandfrisk/black_brown_and_targeted_online.pdf.

[13] Bernard E. Harcourt, *Rethinking Racial Profiling*, 71 U. CHI. L. REV. 1275, 1275-76 (2004).

[14] Jodi M. Brown & Patrick A. Langan, Bur. Just. Statistics, *Policing and Homicide, 1976-1998* (2001).

[15] *See* Center for Policing Equity, *Nation's First Police Profiling Database Awarded Grant By NSF*, Nov. 7, 2013, http://cpe.psych.ucla.edu/images/uploads/database_release_final_%281%29.pdf.

[16] *E.g.*, Adam Liptak, *U.S. Prison Population Dwarfs That of Other Nations*, N.Y. TIMES, Apr. 23, 2008, http://www.nytimes.com/2008/04/23/world/americas/23iht-23prison.12253738.html?pagewanted=all.

[17] PEW CTR. ON THE STATES, ONE IN 100: BEHIND BARS IN AMERICA 2008 (Feb. 2008).

[18] Bonczar, *supra* note 2, at 1.

[19] *Id.* at 2.

[20] *See* Brett E. Garland et al., *Racial Disproportionality in the American Prison Population*, 5 JUST. POL'Y J. 1, 21-26 (2008) (reviewing studies); Alfred Blumstein, *Racial Disproportionality of U.S. Prison Populations Revisited*, 64 U. COLO. L. REV. 743 (1993) (finding that 76% of the black-white incarceration gap stems from arrest patterns).

given arrest offense will contain some weaker or less-serious cases that would not have resulted in arrest on that charge were the arrestee white. If studies of subsequent prosecutorial and judicial decisions do not account for this, they may overlook substantial differences in the cases they are comparing. Controlling for the arrest offense does not result in an apples-to-apples comparison if the arrest offense means something different depending on race.[21]

Researchers have often ignored these problems; instead, they should acknowledge them, and interpret their results cautiously in light of reasonable assumptions guided by the policing literature.[22] For example, if we can assume police probably don't tend to discriminate *against* whites, then estimates of unexplained disparities favoring whites in subsequent processes are likely conservative. Researchers can also offer sensitivity analyses showing effects of competing assumptions about arrest disparities. But the bounds implied by such analyses might be wide, because the uncertainty about arrest disparities is great. To develop a clearer empirical picture of later procedural stages, we need a better handle on policing.

3. Consequences of Criminal Justice Disparities

Data on raw disparities help us to understand who is bearing the burdens of our expansive criminal justice system. These burdens include the social consequences of mass incarceration [23] and the potentially lifelong legal and socioeconomic consequences of having a criminal record.[24] The harms of police interactions are not confined to the guilty—most stops and searches produce no evidence of wrongdoing.[25] Even if no charges are brought, arrest records can produce stigma, job-market consequences, and increased sentences in future cases.[26] And even absent arrest, interacting with police is often stressful and scary, and the experience of being racially targeted can amplify the emotional and dignitary costs.[27]

When police use force, the impacts on individuals and communities can be especially acute, as recent events, including the deaths of Michael Brown and Eric Garner and their aftermath, have amply demonstrated. This Article does not focus primarily on disparities in the use of force, which raise distinct normative concerns

---

[21] The vast majority of sentence-disparity studies compound the problem by using samples of *sentenced* cases and controlling for conviction or sentencing-stage severity measures, failing to account for disparities in charging, plea-bargaining, and sentencing fact-finding. *See* Sonja B. Starr & M. Marit Rehavi, *Mandatory Sentencing and Racial Disparity: Assessing the Role of Prosecutors and the Effects of* Booker, 123 YALE L.J. 2, 39-77 (reviewing literature and explaining this problem).

[22] *See, e.g.*, Sonja B. Starr, *Estimating Gender Disparities in Federal Criminal Cases*, AM. L. & ECON. REV. __ (2014) (forthcoming) (following this approach).

[23] *See, e.g,* Garland et al., *supra* note 20, at 9-14 (reviewing literature).

[24] *See* James Forman, Jr., *Beyond the New Jim Crow*, 87 N.Y.U L. REV. 21, 28-32 (2012).

[25] *See* David A. Harris, *The Stories, the Statistics, and the Law: Why "Driving While Black" Matters*, in RACE, ETHNICITY, AND POLICING, *supra* note 6, at 49.

[26] *See, e.g.*, Gary Fields & John R. Emshwiller, *As Arrest Records Rise, Americans Find Consequences Can Last a Lifetime,* WALL ST. J., Aug. 14, 2014.

[27] *See* Albert W. Alschuler, *Racial Profiling and the Constitution*, 2002 U. CHI. LEGAL F. 163, 212-13; Rod Brunson, *Beyond Stop Rates: Using Qualitative Methods to Examine Racially Biased* Policing, in RACE, ETHNICITY, AND POLICING, *supra* note 6, at 224-33 (interviewing young black men in St. Louis).

and have been hard to study due to lack of data. Rather, although some of the points I make apply to use-of-force disparities, my primary focus (and that of the existing policing-disparity literature) is on investigative interactions, such as stops, searches, and arrests. These issues are, of course, entangled. One key reason those killed by the police are predominantly black men is that black men are vastly more likely to be stopped by the police in the first place, creating far more interactions that can go terribly wrong. And fear of excessive force is one of the factors that can make a "routine" police interaction not just inconvenient, but traumatic.

Of course, policing seeks to prevent crime, and crime damages communities too. And some crimes (for example, homicide) are more commonly committed both by and against people of color[28]—which is unsurprising in light of socioeconomic stratification by race.[29] If the police were to fail to arrest more members of groups that commit more crimes, they might be fairly criticized for *underserving* communities of color, especially because most crimes with victims are intraracial.[30] Such a failure could send an expressive message devaluing victims of color. Analogously, critics have condemned victim-race disparities in capital sentencing for implying that "a black life simply is worth less than a white life."[31]

One might ask, then: do policing disparities harm or benefit the "over-policed" group? In some contexts, the answer is clearly "harm," because the harms are concentrated on the group while the benefits are shared by everyone, or because it is the innocent within the group, not the guilty, who face excessive policing. But in other contexts, there are real tradeoffs between the community's interests in reducing policing burdens and reducing crime, so the answer is less obvious.

These questions resist generalizable answers. Other bodies of empirical scholarship have struggled with estimating the costs of crime and incarceration and the effects of incarceration and policing on crime.[32] The cost of stops, searches, and arrests remain unquantified, as do the effects of community resentment of police. These factors surely vary across groups and localities, and communities' perceptions of what balance of benefits and burdens is appropriate may also vary.[33] Disparity researchers typically do not seek to answer the "net benefits" question (nor do I); they provide a different piece of the puzzle for policymakers.

---

[28] *Crime in the United States 2011, Expanded Homicide Data*, FBI, http://www.fbi.gov/about-us/cjis/ucr/crime-in-the-u.s/2011/crime-in-the-u.s.-2011/offenses-known-to-law-enforcement/expanded/expanded-homicide-data (last visited Oct. 31, 2014).

[29] *E.g.*, Patrick Bayer et al., *Separate When Equal? Racial Inequality and Residential Segregation*, 82 J. URBAN ECON. 32, 32-33 (2014).

[30] *See* RANDALL KENNEDY, RACE, CRIME, AND THE LAW 19 (1997) ("The principal injury suffered by African-Americans in criminal matters is not overenforcement but underenforcement…."); Dan M. Kahan & Tracy L. Meares, *The Coming Crisis of Criminal Procedure*, 86 GEO. L.J. 1153, 1166 (1998); Harris, *supra* note 25, at 49.

[31] Christian Halliburton, *Neither Separate Nor Equal*, 3 SEATTLE J. FOR SOCIAL JUST. 45, 54 (2004).

[32] *See* David S. Abrams, *The Imprisoner's Dilemma: A Cost-Benefit Approach to Incarceration*, 98 IOWA L. REV. 905 (2013) (reviewing this literature).

[33] Within many black communities that once pushed for aggressive law enforcement, public opinion has also shifted over time. WILLIAM J. STUNTZ, THE COLLAPSE OF AMERICAN CRIMINAL JUSTICE 285-87 (2011); Forman, *supra* note 31, at 34-39.

C. Conceptions and Causes of Disparity: Possible Research Objectives

Imagine that policing-disparity researchers could gather whatever data they wanted. What would we prioritize doing with it? This kind of thought experiment is often a valuable guide to empirical research. Here, the answer turns on questions about what kinds of inequality matter.

Consider the following claims about what would constitute racial equality in a city's law enforcement. For the purpose of this exercise and the rest of the Article's hypotheticals, I assume that we are assessing stops. One could substitute other outcomes, such as searches or arrests or downstream outcomes like incarceration, without fundamentally changing the analysis.

1. *The racial distribution of stops should track the racial distribution of the population.*
2. *The racial distribution of stops should track the racial distribution of crime commission.*
3. *A racial disparity in stop rates is only justified if it is proportional to a disparity in crime rates.*
4. *Holding criminal conduct constant, the probability that someone will be stopped should not differ by race.*
5. *The police should not consider race when deciding whom to stop.*
6. *The police should consider race only when doing so improves their odds of catching criminals.*

While some of these formulations sound fairly similar, they express different ideas. Formulation (1) treats raw disparities as troubling inequalities, regardless of the reason. Formulations (2) through (4) focus solely on disentangling the explanatory role of crime (implicitly treating other explanations as unwarranted sources of disparity). Of these, versions (2) and (3) are the most common conceptions of equality tested by the empirical literature; such studies compare stop data to some measure of crime across groups.[34] In Part II, I argue that version (4) is a better way to think about "accounting for crime," and show that it's quite different from versions (2) and (3). Version (4) could be teased out into separate sub-principles requiring racial equality in policing of the innocent and in policing of the guilty.[35]

Formulations (5) and (6) are different in kind. They focus on *decision-making inputs*, not on outcome gaps and their crime justifications. Testing them requires filtering out a broader range of factors (not just crime), to isolate some variety of racial discrimination (a term I use here in its narrow, "disparate treatment" sense). Formulation (5) treats all police uses of race as suspect, while Formulation (6) reflects a recently prominent approach distinguishing between "irrational" prejudice and "rational" use of empirical information.

Many other nuances are possible. When disparity researchers specify an empirical model, they make many decisions about what sources of disparity "count." For

---

[34] *See infra* Part II (reviewing this literature).

[35] A few scholars have called for separate analysis of disparities among the innocent and guilty. David Thacher, *From Racial Profiling to Racial Equality*, unpublished, 8 http://www.ibrarian.net/navon/page.jsp?paperid=1092784&searchTerm=commission+for+racial; R. Richard Banks et al., *Discrimination and Implicit Bias in a Racially Unequal Society*, 94 CAL. L. REV. 1169 (2006); Jeff Dominitz, *How Do the Laws of Probability Constrain Legislative and Judicial Efforts to Stop Racial Profiling?*, 5 AM. L. & ECON. REV. 412, 414 (2003). In Part II, I assess how this can be done effectively.

example, should researchers control for the neighborhood? Some argue that one shouldn't: excessive targeting of minority neighborhoods is a potentially important source of unjustified disparity, not something that should be filtered out.[36] The counterargument is that *intra*-neighborhood disparities—albeit not the whole story—are also important.[37] Filtering out neighborhood effects allows those disparities to be studied in a causally rigorous way, because neighborhoods differ in many ways that could confound analyses of race's effects.

A reader might reasonably object that most empirical papers frame their objectives as descriptive. To describe racial disparities as being explained in part by some variable does not necessarily imply that the part mediated by that variable is normatively *justified*. For example, I may want to disentangle the effects of racial discrimination from those of socioeconomic discrimination even if I believe both are troubling. Empiricists need not specify which explanations are "bad" and which are "good"—they need merely decide which ones are worth disentangling.

This is a fair objection, but it only goes so far. First, many studies *do* explicitly offer estimates of "unwarranted" disparities. Second, empiricists studying race and policing cannot be blind to the context in which their work will be received. The subject matter is high-profile and emotionally and politically charged. Statistics are cited often by virtually everyone in the relevant policy and legal discussions. And those citing statistics clearly do not think of them as being normatively neutral. So when a study reports a bottom-line number that it characterizes as the "racial disparity holding other factors constant," for example, that will be read by most readers as "the unjustified racial disparity." That means researchers should think carefully about what other factors they hold constant.

Just as importantly, authors in that situation should clearly explain their choices and why they made them, what their estimates represent, and what they don't represent. They should not bury these issues deep in the methods section—they should be front and center. Choices like what control variables are included are not technicalities; they define what the study is measuring. Moreover, if researchers aim to tease out the roles of several different factors in producing disparities, they should present results in a way that highlights those roles, rather than just highlighting the disparity that is left over once identified factors are filtered out.[38]

Instead, the empirical literature is largely either quiet or confusing on questions like these. Some studies use methods that do not mask the task they purport to address—for example, stating a concern with "racial profiling," but employing analyses that focus only on ruling out crime differences.[39] Scholars also sometimes

---

[36] *See* Fagan et al., *supra* note 116, at 316.

[37] Indeed, unwarranted intra-neighborhood disparities may raise especially acute fairness concerns: they burden one racial group while neighbors of all races enjoy the crime-prevention benefits. In contrast, extra policing of a neighborhood confers both benefits and burdens on that neighborhood.

[38] Various decomposition methods from labor economics are designed for this purpose, but are rarely used in the criminal justice literature. *See, e.g.*, Starr, *supra*, at 14 tbl.4.

[39] *E.g.*, James E. Lange et al., *Testing the Racial Profiling Hypothesis for Seemingly Disparate Stops on the New Jersey Turnpike*, 22 JUSTICE Q. 193, 194-95 (2005). Some also cite *raw* disparities as evidence of racial profiling. *E.g.*, Cappers, *supra* note 5, at 14-19; Johnson, *supra* note 5, at 1073.

mischaracterize one another's arguments because of confusion about these framing issues—for instance, mistaking the choice to estimate a particular type of disparity for an empirical claim that only that type of disparity exists.[40]

In Parts II and III, I seek to clarify the relationships among different objectives, to discuss what the "right" questions are, and to examine ways of answering them. Part II focuses on the relationship between policing disparities and crime, and Part III focuses on analysis of police decision-making, especially the role of racial discrimination. In highlighting these questions, I do not mean to devalue the importance of measuring raw disparities (version 1). But the raw disparity picture is already relatively clear, so I do not focus on it, instead focusing on *why* the gaps exist.

II. Policing Disparities and Crime

In analyses of the "why" question, the usual starting point is crime. Indeed, it's also often the ending point: many analyses assess *only* whether crime differences can explain policing gaps. This Part examines how crime's explanatory role should be analyzed. I begin with the question of how to measure crime, and then ask whether crime "benchmarks" are being deployed in normatively meaningful ways, even assuming their accuracy. I show that the ubiquitous types of policing-to-crime comparisons do not mean what people imply when they use them. They do not tell us whether people with like conduct are being treated the same way.

A. Crime Benchmarks and the "Denominator Problem"

The search for crime "benchmarks" has long been seen as the key empirical challenge of policing-disparity research.[41] Indeed, the problem of unmeasured crime bedevils essentially all empirical research related to crime.[42] Many police forces have started collecting better data on whom the police are stopping, searching, and arresting. But without data on criminal conduct, we have nothing to compare this to. We have what many call a "denominator problem."[43]

The problem is really twofold. First, criminal justice datasets include no information about cases that never enter the justice system—that is, the vast majority

---

[40] For example, Pickerill et al. claim that legal scholars who focus on racially disparate treatment assume "race is the sole factor that causes police to search motorists. J. Mitchell Pickerill et al., *Search and Seizure, Racial Profiling, and Traffic Stops: A Disparate Impact Framework*, 31 LAW & POL'Y 1, 2 (2009) Engel likewise claims that "the legalistic perspective" assumes away racial differences in crime rates and assumes racial profiling is always ineffective. Robin S. Engel, *A Critique of the 'Outcome Test' in Racial Profiling Research*, 25 JUSTICE Q. 1, 5-9 (2012). Both claim that such legal scholars believe raw disparities are never normatively justified. *Id.* at 9; Pickerill et al. *supra*, at 5. But such claims are not common in legal literature, and none are "legalistic"; the law makes it hard to infer discrimination from disparity. It is perfectly consistent to critique disparate treatment while recognizing that other factors also contribute to disparities. *See, e.g.*, KENNEDY, *supra* note 30, at 150-51.

[41] *E.g.*, Robin Shepard Engel & Jennifer M. Calnon, *Comparing Benchmark Methodologies for Police-Citizen Contacts*, 7 POLICE QUARTERLY 97, 98, 100 (2004); Ridgeway, *supra*, at 19.

[42] *See, e.g.*, Albert D. Biderman & Albert J. Reiss, Jr., *On Exploring the 'Dark Figure' of Crime*, 374 ANNALS OF THE AM. ACAD. OF POLIT. & SOCIAL SCIENCE 1, 1 (1976).

[43] *E.g.*, Meaghan Paulhaumas et al., *State of the Science in Racial Profiling Research*, in RACE, ETHNICITY, AND POLICING, *supra* note 6, at 239.

of crimes. For example, surveys suggest there are at least hundreds of millions of drug crimes per year, but only about 1.5 million drug arrests.[44] Even most reported violent and property crimes go unsolved.[45] Second, even for those who *do* interact with police, we lack objective conduct measures. Rather, we have what officers write down, and in a study of policing disparity, one shouldn't assume this is accurate.

This situation is not really a data collection failure, however: nobody thinks the data in question *ought* to be collected comprehensively. The value that society places on freedom from constant surveillance requires that the vast majority of crimes go uncounted (not to mention unpunished). So researchers must rely on imperfect, incomplete proxies. But what makes a good proxy?

Some scholars use arrest rates, often from a prior time period, to stand in for crime rates. But this introduces a troubling circularity: arrests are discretionary decisions by the very actors whose stops are being studied. What is really being asked is thus not "Does crime explain stop disparities?" but "Are stop disparities bigger than arrest disparities?" Arguably, arrest benchmarks might sometimes be defensible as deliberately conservative,[46] but they should generally be avoided.

Reported crime is a better alternative, but is still imperfect. Crime reports are collected by local agencies and compiled by the FBI.[47] The principal problem is that most crime goes unreported—about half of violent crimes and 60% of property crimes, according to victim surveys.[48] Moreover, minor and victimless crimes are

---

[44] *See* Office of Nat'l Drug Control Pol'y, Fact Sheet, http://www.whitehouse.gov/sites/default/files/ondcp/Fact_Sheets/ nsduh_fact_sheet_9-7-11_0.pdf (about 23 million Americans use illicit drugs each month). For arrest rates, see the online calculation tool at http://www.bjs.gov/index.cfm?ty=datool&surl=/arrests/index.cfm#. *See also Impaired Driving*, CDC, http://www.cdc.gov/motorvehiclesafety/impaired_driving/impaired-drv_factsheet.html (finding 1% arrested out of 112 million self-reported drunk-driving instances).

[45] Nationally, under 20% of reported property crimes and 45% of reported violent crimes are cleared. *Crime in the United States 2012, Clearances*, FBI, http://www.fbi.gov/about-us/cjis/ucr/crime-in-the-u.s/2012/crime-in-the-u.s.-2012/offenses-known-to-law-enforcement/clearances.

[46] For instance, Fagan et al., *supra* note 116, at 318-19, and Andrew Gelman et al., *An Analysis of the NYPD's Stop-and-Frisk Policy in the Context of Claims of Racial Bias*, 102 J. AM. STAT. ASSOC. 813, 818-20 (2012), find disparities in stop-and-frisk rates after controlling for prior-year arrests. Their estimates are likely conservative, assuming arrest and stop disparities cut in the same direction.

[47] The FBI's Uniform Crime Reports (UCR) provide summary data for certain crimes, but have no race information except for homicides. *Crime in the United States 2012,* http://www.fbi.gov/about-us/cjis/ucr/crime-in-the-u.s/2012/crime-in-the-u.s.-2012/offenses-known-to-law-enforcement.
Some agencies participate in the National Incident-Based Reporting System, which includes suspect race. U.S. DEP'T OF JUSTICE, FBI, NATIONAL INCIDENT-BASED REPORTING SYSTEM: DATA COLLECTION GUIDELINES (2000), http://www2.fbi.gov/ucr/nibrs/manuals/v1all.pdf. Some studies use crime-report data from local sources. *E.g.*, Howard P. Greenwald, *Final Report: Vehicle Stops in Sacramento, California,* Report to the City of Sacramento (2001); Ridgeway, *supra*, at 13.

[48] Jennifer L. Truman, *Criminal Victimization, 2010*, Bureau of Justice Statistics, Bulletin, September 2011, 1, http://www.bjs.gov/content/pub/pdf/cv10.pdf. Garland et al., *supra*, at 19-20, respond that reported crime is the right comparison point because "the criminal justice process does not begin until the police become aware of a crime," so police are not responsible for reporting disparities. But much policing is proactive, not report-driven. *See* Ridgeway, *supra*, at 18 (finding 30% of NYPD stops responded to citizen calls). Regardless, here, we are assessing whether policing burdens are fairly distributed conditional on criminal conduct, not whether to blame police for any maldistribution.

almost never reported,[49] so reported-crime benchmarks are usually based on "index crimes,"[50] violent crimes, or even just homicide.[51] But if stop or arrest rates are driven heavily by drugs or minor crimes, comparing them to such benchmarks may introduce bias. Racial differences in crime rates are believed to be far greater for violent crime, especially homicide.[52] So using violent-crime benchmarks to proxy for *all* crime risks substantially overstating the extent to which disparities in stops or arrests are explained by differences in criminal conduct.

Another alternative is survey data. Some surveys, such as the Census Bureau's National Crime Victimization Survey (NCVS), gather data about crimes with victims.[53] Other surveys ask individuals to self-report drug use. These generally show only minor racial differences in drug use,[54] findings that scholars often contrast with large race gaps in drug arrest rates.[55] Survey benchmarks avoid some of the concerns raised above: they cover crimes not reported to police, are not themselves shaped by police, and encompass a wider variety of crimes. But there remain concerns about accuracy.[56] Also, national surveys' samples are not designed to produce valid *local* estimates.[57] And most drug surveys cover only use, not supply.[58]

---

[49] Robert J. Sampson & Janet L. Lauritsen, *Racial and Ethnic Disparities in Crime and Criminal Justice in the United States*, 21 CRIME & JUST. 311, 317 (1997).

[50] The eight crimes used in the Uniform Crime Reports "index": murder, forcible rape, arson, larceny, robbery, burglary, car theft, and aggravated assault.

[51] *E.g.*, Fagan et al., *supra* note 116, at 318-19 (using homicide); Greg Ridgeway, *Analysis of Racial Disparities in the New York Police Department's Stop, Question, and Frisk Practices*, RAND Corp. Rep't TR-534 xii, 19, http://www.rand.org/pubs/technical_reports/TR534.html (using violent crime). Again, this choice might be defensible as conservative: Fagan et al. find disparities despite using the homicide benchmark, which is striking given the likely downward-biasing effect. But it is clearly inappropriate to use violent-crime benchmarks to show *lack* of disparity, as Ridgeway does.

[52] Richard S. Frase, *What Explains Persistent Racial Disproportionality in Minnesota's Prison and Jail Populations*, 38 CRIME & JUST. 201, 238 (2009).

[53] For the questionnaire, see http://www.bjs.gov/content/pub/pdf/ncvs2_2012.pdf. *See* Shima Baradaran, *Race, Prediction, and Discretion*, 81 GEO. WASH. L. REV. 157, 201 (2013) (citing the NCVS).

[54] Monitoring the Future I, 9, http://monitoringthefuture.org//pubs/monographs/mtf-vol1_2013.pdf; Dep't Health & Hum. Servs., *Results from the 2013 Nat'l Survey of Drug Use and Health* 26, http://www.samhsa.gov/data/sites/default/files/NSDUHresultsPDFWHTML2013/Web/NSDUHresults2013.pdf.

[55] *E.g.,* MICHELLE ALEXANDER, THE NEW JIM CROW 6-7; Yonette F. Thomas, *The Social Epidemiology of Drug Abuse*, 32 AM. J. PREVENTATIVE MED. S141 (2007); *see also* Gross & Barnes, *supra* note 11, at 681 (comparing survey data to traffic stop rates); Christopher L. Griffin, Jr. et. al., *Corrections for Racial Disparities in Law Enforcement*, 55 WM. & MARY L. REV. 1365, 1381-82 (2014) (using surveys on DWIs).

[56] *See* Barry Spunt, *Self-Report Surveys*, in 4 ENCYCLOPEDIA OF CRIME AND PUNISHMENT 1465, 1466-67 (David Levinson ed.) (2002); Arthur H. Garrison, *Disproportionate Minority Arrest*, 23 NEW ENG. J. ON CRIM. & CIV. CONFINEMENT 29, 42-45 (1997).

[57] Brian Wiersema, *Area-Identified National Crime Victimization Survey Data*, NCVOR Census Ctr. Tech. Paper 1 (1999), http://www.ncovr.heinz.cmu.edu/docs/Wiersema-Area-Identified%20NCVS.pdf; Dep't Health and Human Servs., Substance Abuse and Mental Health Admin., *State Estimates of Substance Abuse from the 2006-07 National Surveys on Drug Use and Health* 7-9 (2009).

[58] One survey that did ask questions about drug sales found somewhat higher rates among blacks, albeit not high enough to explain the arrest gap. *See* Frase, *supra* note 52, at 239.

Given existing datasets' limits, some researchers have gathered new data by physically observing public behavior—usually traffic violations. The seminal example was John Lamberth's 1994 study of the New Jersey Turnpike, in which researchers drove just over the speed limit and observed the drivers who passed them; the racial composition of speeders was compared to Turnpike stop data.[59] Subsequent highway speeding studies have used radar.[60] Alpert et al. similarly physically observed traffic violations on city streets.[61] Outside the traffic context, Dabney et al. used video cameras in a store to observe shoplifting—an innovative method, though the study's analysis is unfortunately hard to interpret.[62] The study also illustrates a key concern with benchmarking generally: the appropriateness of comparing behavioral and policing-outcome data from different sources. The authors compare shoplifting apprehension figures from a national study to shoplifting rates at a single store, which could be atypical. The highway-speeding studies do better on this score, comparing observed behavior with stop data from the same highways.

The direct-observation method has promise for behaviors (like traffic offenses) occurring in predictable, public locations where researchers or recording devices can be stationed. It could usefully be applied to law enforcement checkpoints, which could record agents' behavior and that of individuals passing through, plus any computer-database information that agents observe. This approach could go beyond traditional benchmarking, measuring individuals' behavior and agents' treatment for the same sample, instead of comparing across different datasets.[63] But most policing contexts are not a good fit for direct observation—most crime occurs in private or unpredictable places. For such conduct, the best available benchmark probably remains victim reports or survey data, despite these sources' limitations.

B. Criminal Conduct as a Possible Justification for Policing Disparities

Once one has settled on a crime benchmark, what should one do with it? To answer this question, we need to ask why so much research focuses on whether crime explains policing disparities to the exclusion of other possible explanations.

---

[59] John Lamberth, *Revised Statistical Analysis of the Incidence of Police Stops and Arrests of Black Drivers/Travelers on the New Jersey Turnpike Between Exits or Interchanges 1 and 3 from the Years 1998 through 1991*, Nov. 11, 1994. Fifteen percent of speeders and 35% of those stopped were black.

[60] *See* Lange et al., *supra* note 39, at 211-12; Robin S. Engel et al., Pennsylvania State Police Project on Police-Citizen Contacts, Year 2 Report 64-65, 110 (2002)

[61] Geoffrey P. Alpert et al., *Investigating Racial Profiling by the Miami-Dade Police Department*, 6 CRIMINOLOGY & PUB. POL'Y 25, 36, 41-44 (2007) (finding no unjustified racial disparity in stops).

[62] The authors find no racial disparities in shoplifting after controlling for shopper behaviors such as product-tampering, but these controls seem to filter out part of the shoplifting conduct itself. Dean A. Dabney et al., *Who Actually Steals? A Study of Covertly Observed Shoplifters*, 21 JUST. Q. 693, 711 (2004).

[63] Federal agencies have already conducted self-studies designed to produce racial benchmarks for comparisons to checkpoint stops. Bureau of Justice Statistics, Assessing Measurement Techniques for Identifying Race, Ethnicity, and Gender N.C.J. Report 196855, http://www.bjs.gov/content/pub/ascii/amtireg.txt (2003). However, these studies merely recorded the race of those passing through, plus (in an airport security study) some additional information such as gender, age, and number of carry-ons. This could easily be extended to record agent responses, individual behavior, and computer-system information.

One possibility is crime's descriptive importance. Many scholars have argued that crime differences are the single most important explanation for race gaps in U.S. arrest and incarceration rates.[64] But this does not really explain why so many analyses focus on crime exclusively. Some policing-to-crime comparisons (such as the common comparison of drug arrest rates to use rates) do *not* actually find that crime substantially explains policing disparities. And even when it does, other factors may also be important. Indeed, when other contributors to disparity are layered on top of substantial crime differences, it amplifies their consequences.[65] In any event, crime-disproportionality analyses are not designed to support strong causal inferences. They tell us whether policing rates are in line with what we would expect based on crime patterns, not why they are or are not.[66]

But there is another reason for the focus on crime. Many researchers and commentators appear to treat crime differences as potential *justifications* for disparities in police interactions, not just explanations. Meanwhile, policing disparities that are disproportionate to crime differences are presented, explicitly or by implication, as unjustified. The crime-disproportionality question seems to get at many people's core intuitions about what makes racial disparities in the justice system not just unfortunate, but unfair. So what underlies these intuitions?

Start with an oft-repeated principle: like cases should be treated alike. The limits of this principle have been much debated,[67] but it is clearly a widely shared intuition, at least as a default rule.[68] A sensible way of understanding the literature's emphasis on crime benchmarks is that criminality is what determines which cases are meaningfully "like." If so, when we say policing should be "proportionate" to crime, we imply a specific objective: *equal policing of people of different racial groups conditional on their criminal conduct*. Assuming a simplified binary guilty/innocent division, this objective requires that (1) the probability that an innocent person will be stopped does not vary by race, and (2) the probability that a guilty person will be stopped does not vary by race. We can express these objectives as follows:

$$(1)\ D_I = \frac{S_{I1}/P_{I1}}{S_{I2}/P_{I2}} = 1, \text{ and } (2)\ D_G = \frac{S_{G1}/P_{G1}}{S_{G2}/P_{G2}} = 1$$

[64] *E.g.*, KENNEDY, *supra* note 30, at 23, TONRY, *supra* note 73, at 79; Forman, *supra* note 31, at 31-32.

[65] For example, suppose crime could explain the vast majority of the 6-to-1 black-white incarceration gap, i.e., whites were only 10% less likely to be incarcerated than blacks with the same criminal conduct. Even then, that gap would be consequential: reducing black incarceration by 10% would mean restoring the liberty of more than 1% of all black men under 35 in the United States (1 in 9 of whom are now in prison). *See* M. Marit Rehavi & Sonja B. Starr, *Racial Disparity in Federal Criminal Sentences*, 122 J. POLIT. ECON. 1320, 1349-50 (2014) (making a similar calculation).

[66] For instance, policing patterns that are unexplained by crime might result from racial discrimination, or from applying race-neutral criteria that are differentially accurate across races.

[67] *E.g.*, David A. Strauss, *Must Like Cases Be Treated Alike?*, U. Chi. Pub. L. & Legal Theory Working Paper No. 24, at 3 (2002); Kenneth Winston, *On Treating Like Cases Alike*, 62 CAL. L. REV. 1 (1974).

[68] *See* Strauss, *supra*, at 3 (arguing, however, that there are often good reasons to depart from it).

S stands for "stops" and P for "population," subscripts *I* and *G* stand for "innocent" and "guilty," and subscripts 1 and 2 identify the two racial groups being compared. D_I and D_G are ratios of the groups' stop rates among the innocent and guilty, respectively. One can easily complicate proposition (2) to accommodate many varieties of criminality; the objective remains that those with the same culpability should face the same probability of apprehension.

Why does criminal behavior determine "likeness"? One strong argument is based on moral desert. Policing is not punishment, but it facilitates it, so if we think criminals deserve punishment, we should want them to face police stops, searches, and arrests; we should also want the innocent to avoid such interactions. The police lack perfect information, so they cannot stop every criminal and will stop some innocents. But if the probability of these errors varies by race, there is racial inequality unjustified by criminal conduct.[69]

Two clarifications are important here. First, I am not suggesting that the higher-crime group *as a whole* deserves heavier policing. The concept of "group desert" finds no support in retributive precepts,[70] and affixing blame to an entire race would be particularly repugnant. Nonetheless, targeting of guilty *individuals* can in the aggregate produce group outcome differences. Second, I exclude use of force from the discussion of whether crime differences "justify" policing disparities. I do not think that criminal culpability means that one deserves to be subjected to physical violence by police. Use of force may occasionally be justified based on different kinds of moral considerations, focused on imminent risk. But such situations cannot be identified using crime benchmark data.

An alternative to the retributive rationale for treating crime as a special justification is to focus on the rule of law. The "like cases" principle is often described as a rule-of-law value: it "entails and is entailed by conformity to law."[71] If so, then law determines what cases are "like"—in the policing context, criminal law specifically. This argument focuses more on preventing arbitrary decision-making than on outcome fairness, but still implies that policing rates should be equal across racial groups conditional on criminal conduct.[72]

Note that one can embrace this objective without thinking that we shouldn't worry about disparities that *are* explained by criminal conduct. But in that case, the worry wouldn't be that policing is misaligned with what it's supposed to target.

---

[69] *See* Thacher, *supra* note 35, at 2 (arguing for goal of "racial equality within morally homogeneous groups").

[70] *See, e.g.*, Neal Kumar Katyal, *Conspiracy Theory*, 112 YALE L.J. 1307, 1369 (2003); Joshua Dressler, *Reassessing the Theoretical Underpinnings of Accomplice Liability*, 37 HASTINGS L.J. 91, 103-08 (1985).

[71] Winston, *supra*, at 5.

[72] A counterargument is that perhaps what should determine "likeness" is behavior *outwardly signaling* likely criminality, legally justifying a stop. This would be inconsistent with the moral desert objective, but one could defend it from the rule-of-law perspective. We lack data on underlying suspiciousness, however, and it is a readily manipulable characterization. The problems I identify below with policing-to-crime comparisons assume that actual criminality is the intended measure of "likeness." If suspiciousness were the right measure and if crime is a poor proxy for it, that would raise additional problems for policing-to-crime comparisons, since they have the wrong comparison point entirely.

Instead, it might focus our attention on root causes of crime differences (for example, educational inequities and poverty), or else on the overall scope and severity of our criminal law and its enforcement, given its disparate impact.[73] Those who present policing-to-crime comparisons do not necessarily imply that these other types of inequality don't matter, but they do imply that inequality in police interactions among people who have done the same thing is a distinct fairness problem that merits empirical measurement and policy concern.

Richard Banks, Jennifer Eberhardt, and Lee Ross have called for separate attention to disparities in policing of the guilty and of the innocent, respectively. This call is persuasive: stops, searches, and arrests of the innocent are different kinds of events from the same treatment of the guilty, and affect communities differently. The authors go farther, however, asserting that when offending rates differ, "one cannot attain equality across groups with respect to both the investigation of the innocent and the apprehension of the guilty."[74] They claim:

> If the crime rate is higher among Blacks than Whites, but the rates of investigation are the same, then an African American criminal will be less likely to be apprehended than a white criminal. To equalize across groups the likelihood that a criminal will be apprehended would require increasing stop-search rates among Blacks, which would have the unfortunate consequence of also increasing the likelihood that innocent Blacks are investigated…. Either innocent members of the higher crime rate group will be subject to a greater likelihood of investigation, or a greater percentage of criminals from the higher crime rate group will be permitted to stay at large.[75]

This dilemma sounds daunting, and indeed, tradeoffs like this can sometimes emerge in practice, and are worth highlighting. But the authors' claim is much too strong. It is certainly possible to simultaneously equalize across races stop probabilities for both the innocent and the guilty, even with very different crime rates. For instance, randomized enforcement, such as at some DUI checkpoints, would always satisfy both requirements. As another example, suppose police stop suspected drunk drivers if and only if those drivers swerve. Suppose that regardless of race, 60% of drunk drivers swerve and 10% of sober drivers swerve. The race-neutral swerving criterion means that 60% of drunk drivers get stopped and 10% of sober drivers get stopped, regardless of race, whatever each racial group's drunk-driving rate is. Stop *productivity* (the percentage of stops that catch drunks, which is not the same as the percentage of drunks who get stopped) will be higher for whichever group has a higher drunk-driving rate. But this is troubling only if equal productivity is an important metric of equality. As I argue below, it isn't.[76] Many other examples are given in the next Section and the appendix.

---

[73] *See, e.g.*, MICHAEL TONRY, MALIGN NEGLECT: RACE, CRIME, AND PUNISHMENT IN AMERICA 79-80, 105 (1995); Kenneth B. Nunn, *The "Darden Dilemma": Should African Americans Prosecute Crimes?*, 68 FORDHAM L. REV. 1473, 1487-89 (2000).

[74] Banks et al., *supra* note 35, at 1178-79.

[75] *Id.* at 1179.

[76] Banks et al. also seem to conflate the question whether black and white criminals face the same rate of apprehension with the question whether the police leave "more Black criminals than White criminals at large as a proportion of the group's population." *Id.* at 1179. Equalizing the prevalence of

So there is nothing necessarily incompatible about racial equality in policing of both the guilty and the innocent—and if we want to assess whether criminal conduct "justifies" policing differences, we should ask whether those objectives are being achieved. Unfortunately, as I show below, the literature comparing policing rates to crime rates has overwhelmingly failed to ask that question.

C. Disproportionality Ratios: Are We Asking the Right Questions?

When researchers and commentators ask whether policing is "proportional" to crime, what proportion are they referring to? Two comparison types are widespread. I show here that neither is consistent with the "treat like cases alike" objective. Again, my hypotheticals use "stops" as the policing outcome of interest, but others could substitute; many examples that I cite from the literature focus on arrests.

First, many analyses ask: Is the ratio of police interactions for two groups different from the ratio of crimes? For example, Prof. Michael Tonry writes that for adult men, the ratio of black to white self-reported violent crimes is "4:1, which is 'very similar to differences observed in the *Uniform Crime Reports* of arrests for violent offenses at this age,' unlike the adolescent years when the self-report ratio is 1.5:1 and the arrest ratio is 4:1."[77] His implication is that among adults, the arrest gap is explained by offending differences, but among adolescents, it's not.

The implied conception of inequality is sometimes formally expressed by dividing the ratio of stop rates by the ratio of crime rates.[78] I label this "Stop Ratio/Crime Ratio" or "Ratio/Ratio" disproportionality:

$$Stop\ Ratio/Crime\ Ratio = \frac{Stop\ Rate\ 1 / Stop\ Rate\ 2}{Crime\ Rate\ 1 / Crime\ Rate\ 2} = \frac{S_1/S_2}{P_{G1}/P_{G2}}$$

For racial groups 1 and 2, S is the number of stops and P_G is the number of criminals (or crimes[79]). These comparisons are sometimes framed in terms of *rates* of stops and

---

criminals-at-large is not the same as equalizing apprehension rates conditional on criminality. The former is an appealing aspiration, but may be unachievable via policing alone if underlying crime rates are quite different, unless crime rates are highly elastic. Even if it could be done, it might require a police presence that communities would find intolerable—and would likely require one group to have a much higher apprehension rate for both the innocent *and* the guilty than the other group has. Realistically, policing can't cancel out the many social and historical causes of crime-rate differences.

[77] TONRY, *supra* note 73, at 78-79. For similar examples, see Leadership Conference on Civil Rights & Leadership Conference Education Fund, *Justice On Trial: Racial Disparities in the American Criminal Justice System* (2000), http://www.civilrights.org/publications/justice-on-trial/juvenile.html (citing the same ratios as Tonry's concerning adolescents); Baradaran, *supra* note 53, at 201-02 (pointing to relatively similar ratios as evidence of lack of substantial unjustified disparity); American Civil Liberties Union, *Blacks Found to Be 3.3 Times More Likely to Be Arrested for Marijuana Possession Than Whites in Michigan, Despite Equal Usage Rates* (Jun. 4, 2013), https://www.aclu.org/criminal-law-reform/blacks-found-be-33-times-more-likely-be-arrested-marijuana-possession-whites (pointing to different arrest and crime ratios as evidence of unjustified disparity); Maia Szalavitz, *Study: Whites More Likely to Abuse Drugs Than Blacks*, TIME, Nov. 7, 2011 (same); Halliburton, *supra*, at 55 (same); Griffin, *supra*, at 1381-82 (comparing arrest and self-report rates for DWI).

[78] *See, e.g.*, Robert D. Crutchfield, *Warranted Disparity? Questioning the Justification of Racial Disparity in Criminal Justice Processing*, 36 COLUM. HUM. RTS. L. REV. 15, 30 tbl. 2 (2004).

crimes, but it is equivalent to use total numbers instead; the "per capita" parts of each rate term cancel.[80] The implication is that when crime differences explain policing disparities, the stop ratio equals the crime ratio (so the Ratio/Ratio measure is 1). "Guilt" and "innocence" are generally presented as binary, but one could extend the concept to more complex scenarios.[81]

The other common framing compares a group's share of total stops to its share of total crime. I refer to this as "Stop Share/Crime Share" or "Share/Share" disproportionality. For example, the debate over stops on the New Jersey Turnpike has revolved around these comparisons, beginning with Lamberth's finding that the black stop share was 35% while the black share of speeders was 15%. On the other side, prominent commentator Heather MacDonald cites a finding that 25% of the drivers going 80mph in a 65mph zone were black; she then states: "Blacks are actually stopped less than their speeding behavior would predict—they are 23 percent of those stopped."[82] Similarly, a RAND Corporation study of NYPD's stop-and-frisk policy (cited heavily by the NYPD) concludes that "blacks are substantially understopped" because they constitute 53% of all stops but 69% of violent-crime suspect descriptions.[83] Other examples abound.[84]

These comparisons imply that racially equitable policing should result in a racial distribution of stops that parallels the racial distribution of crime. So if a group's stop share is bigger than its crime share, it is overpoliced relative to other groups, after accounting for crime; if the stop share is smaller, it is underpoliced; if they are

---

[79] The denominator here refers to guilty *people*, which tracks some prominent studies—for example, those that count speeding drivers. Alternatively, one could use *number of crimes* in a given period. For this purpose, these approaches are conceptually equivalent: the ratio of total crimes in a period should equal the *average* ratio of people committing a crime at all moments during that period.

[80] *See, e.g.*, Baradaran, *supra* note 53, at 201-02 (using raw numbers instead). Using raw numbers changes the stop ratio and the crime ratio, but in the same proportion, making the ratio-of-ratios equivalent. If the stops and crime data come from different samples with different racial compositions, however, the rates on the left must first be calculated within each sample.

[81] For example, one could imagine a denominator combining various crime frequencies into some severity-weighted measure, or calculate separate ratios for each crime condition.

[82] Heather Mac Donald, *The Racial Profiling Myth Debunked*, CITY JOURNAL (2002), http://www.city-journal.org/html/12_2_the_racial_profiling.html.

[83] Ridgeway, *supra*, at 19.

[84] *E.g.*, Floyd v. City of New York, 959 F. Supp. 2d 540, 584 (S.D.N.Y. 2013) (citing another NYPD share/share comparison); REBECCA M. BLANK ET AL., MEASURING DISCRIMINATION 193 (2004) (comparing 18% black speeding share to 73% search share); KATHERYN K. RUSSELL, THE COLOR OF CRIME 35 (1998) (comparing 38% black crack-user share to 85% conviction share); Robert J. Sampson & Janet L. Lauritsen, *Racial and Ethnic Disparities in Crime and Criminal Justice in the United States*, 21 CRIME & JUSTICE 311, 328 (1997) (comparing 56% black robbery suspect share to 61% arrest share); Joan Zorza, *Mandatory Arrest for Domestic Violence Why It May Prove the Best First Step in Curbing Repeat Abuse*, 10 CRIM. JUST. 2, 52 (1995) (comparing shares of domestic violence reports and arrests); Tracey Maclin, *Race and the Fourth Amendment*, 51 VAND. L. REV. 333 (1998) (comparing speeding and stop shares); Lange et al, supra note 39, at 211; Stacey Patton, *If You're White, That Joint Probably Won't Lead to Jail Time*, WASH. POST, Jan. 12, 2014 (comparing 14% black drug-user share to 34% drug-arrest and 53% drug-incarceration shares); Heather MacDonald, *How to Increase the Crime Rate Nationwide*, WALL STREET J., June 12, 2013, A17 ("Blacks, at 55% of all police-stop subjects in 2012, are actually understopped compared with their 66% representation among violent criminals.").

about equal, there is no unjustified racial disparity. In the empirical literature, such comparisons are often formalized as another kind of disproportionality ratio:

$$Stop\ Share/Crime\ Share = \frac{S_1/S}{P_{G1}/P_G}$$

S_1 and P_{G1} represent stops and criminals (or crimes) in Group 1 and S and P_G represent stops and criminals (or crimes) in the whole population.[85]

These two disproportionality measures are not the same. The Ratio/Ratio measure compares two groups; the Share/Share measure compares one group to everybody. One obvious distinction is that "everybody" may include more than two groups. But even assuming a two-group world, the comparisons differ, because the Share/Share approach compares Group 1 to the whole population *including itself*, while the Ratio/Ratio measure compares it to Group 2 only. The measures always cut in the same direction, but the Share/Share measure is always closer to 1; it is "diluted" because it is partially a self-comparison.[86] Another difference (illustrated in the Appendix) is that the Share/Share measure depends on the groups' sizes.

More importantly, neither the Ratio/Ratio comparison nor the Share/Share comparison tests whether there are racial disparities in policing of people with the same criminal conduct. These comparisons can easily be misinterpreted as answers to that question (indeed, they certainly seem *intended* to answer it), but this interpretation is misleading. I illustrate this point here with just two examples. The Appendix provides many more, plus proofs of the key propositions.

The core problem can be explained fairly simply. Suppose two racial groups have different crime rates, and neither the probability of an innocent being stopped nor that of a criminal being stopped varies by race. Should we expect the racial distribution of stops to parallel the distribution of crimes? Our intuitions may say yes, but the answer is no—not unless only criminals are stopped.[87] When we include the innocent in the numerators of all the ratios (the stop terms) but not the denominators (the crime terms), the comparisons don't work. The effect of this

---

[85] *E.g.*, Engel et al., *supra* note 60, at 104-09 (calculating traffic stop share/speeding share ratios for each of 27 Pennsylvania counties.); *see* Engel, *supra* note 40, at 9 (describing wide use of this method).

[86] See Appendix, Proof 2.

[87] Although this problem has been widely overlooked, the district court in the recent *Floyd* stop-and-frisk litigation in effect identified an extreme example of it, saying that comparisons to crime shares were irrelevant because almost all those stopped were innocent. 959 F.Supp.2d at 584-85. The court held that the better benchmark was each racial group's *overall* population (within neighborhoods), not its guilty population. *Id.* (It relied on analyses that included other controls as well.) Note, however, that this is alternative will usually raise the opposite problem—not "accounting for crime" at all, instead of overcorrecting. We would only expect the distribution of stops to mirror the population distribution if the police stop people at random. That may have been nearly true in *Floyd*; only 2% of those stopped had any kind of contraband, which the plaintiffs argued made it essentially like a random stop. *Floyd v. City of New York,* 08 Civ. 1034, Report of Jeffrey Fagan 63-65. Still, in most contexts, stop rates are surely at least somewhat higher for the guilty than the innocent. In the next Section, I discuss alternative methods that neither assume that everyone stopped is guilty *nor* that those stopped are a random subset of their racial groups.

numerator/denominator mismatch cuts in a specific direction: it inflates the stop-crime ratio more for the lower-crime group than for the higher-crime group. So it looks like the lower-crime group is relatively overstopped after "accounting for crime," even though this hypothetical assumes equitable policing.

In practice, it's rarely only criminals who are stopped, searched, or arrested, even if the police never overreach their authority—the law doesn't expect perfect accuracy in policing. Indeed, we can expect *many* police interactions with the innocent even when police are quite good at predicting guilt. That's because in most contexts, the innocent far outnumber the guilty. For example, if stop rates among the guilty are 20 times those among the innocent, but there are 20 times as many innocent people, half of those stopped will be innocent.

Let's work through two examples. Both assume a simple world with two racial groups (black and white) and two criminal conduct conditions (innocent and guilty). I also assume the premise of defenders of policing disparities is correct: black crime rates are higher. I'll show that it's precisely when crime-rate disparities are large that Share/Share and Ratio/Ratio comparisons are most misleading.

The first example assumes racially equitable policing conditional on criminal conduct. Suppose the police, looking for weapons, stop pedestrians if and only if they appear to have objects in their jacket pockets. Assume that regardless of race, 50% of weapons-carriers ("guilty") and 25% of non-carriers ("innocent") have bulging pockets and are stopped—but black pedestrians are twice as likely to have weapons (40% versus 20%). Table 1a gives the expected breakdown of stops if the police encounter 100 black and 100 white pedestrians, and Table 1b calculates the Ratio/Ratio and Share/Share measures.

Table 1a. Example Assuming Racial Equality Conditional on Conduct

Both Races: Assume Innocent Stop Rate = 25%, Guilty Stop Rate = 50%

P: 100 Black, 100 White	Innocents *[P_I]*	Innocent Stops *[$S_I = P_I * 0.25$]*	Guilty *[P_G]*	Guilty Stops *[$S_G = P_G * 0.5$]*	Total Stops *[$S = S_I + S_G$]*
Black (*40% Guilty)*	60	15	40	20	35
White (*20% Guilty)*	80	20	20	10	30

Table 1b. Ratio/Ratio and Share/Share Calculations: Same Example

	Black/White Ratio	Black Share of Total
Total Stops	35/30 = 1.167	35/65 = 0.538
Guilty	40/20 = 2	40/60= 0.667
Disproportionality	*Stop Ratio/Crime Ratio* = 1.167/2 = 0.583	*Stop Share/Crime Share* = 0.538/0.667 = 0.807

Both disproportionality measures are well below 1, from which scholars and commentators would typically infer that black pedestrians are "understopped" once you "account for crime." For example, the Ratio/Ratio interpretation would be that after accounting for crime, black pedestrians are 58% as likely to be stopped as white pedestrians. The Share/Share comparison meanwhile implies that the black stop

share is 81% of what it "should" be. Both are very misleading, given that we've assumed equal policing conditional on criminal conduct. (The Share/Share measure is less spectacularly wrong here, but it isn't always; it's just always closer to 1.)

Now consider a second example, in which policing *isn't* racially equitable. Table 2a makes the same assumptions as in Table 1a except for lower white stop probabilities. Innocent white pedestrians get stopped at a rate of just 15%, while innocent black pedestrians get stopped at a 25% rate. Meanwhile, guilty white pedestrians get stopped at a 30% rate, and guilty black pedestrians at a 50% rate. Table 2b shows the resulting Ratio/Ratio and Share/Share calculations.

Table 2a. Example Assuming Racial Disparity Conditional on Conduct

Black: Assume Innocent Stop Rate = 25%, Guilty Stop Rate = 50%
White: Assume Innocent Stop Rate = 15%, Guilty Stop Rate = 30%

P: 100 Black, 100 White	Innocents *[P_I]*	Innocent Stops *[S_I]*	Guilty *[P_G]*	Guilty Stops *[S_G]*	Total Stops *[$S=S_I + S_G$]*
Black (*40% Guilty)*	60	15	40	20	35
White (*20% Guilty)*	80	12	20	6	18

Table 2b. Ratio/Ratio and Share/Share Calculations: Same Example

	Black/White Ratio	Black Share of Total
Total Stops	35/18 = 1.944	35/53 = 0.660
Guilty	40/20 = 2	40/60= 0.667
Disproportionality	*Stop Ratio/Crime Ratio* = 1.944/2 = 0.972	*Stop Share/Crime Share* = 0.660/0.667 = 0.991

Here, the racial distribution of stops approximately equals the racial distribution of crimes. The black stop share is 66% and the black crime share is 66.7%; the crime ratio also slightly exceeds the stop ratio. These are the kinds of numbers that police departments and their supporters routinely cite to show that differences in stop rates are fully explained by crime differences. Indeed, if anything, these comparisons create the impression that black pedestrians are very slightly *understopped*; both disproportionality measures are slightly below 1. But that conclusion would be very misleading. Remember, this example assumes that the police are substantially *overstopping* black pedestrians—they are 1.67 times as likely to stop a black pedestrian than a white pedestrian with the same criminal conduct.

These problems with Share/Share and Ratio/Ratio comparisons aren't an artifact of the particular numbers I chose. The Appendix's proofs and additional examples show that anytime group crime rates differ and some innocents are stopped, the comparisons are at least somewhat misleading, and under many realistic sets of facts, they can be drastically so. Specifically:

- Racially equal policing of both the innocent and the guilty always results in the higher-crime group having a lower stop/crime ratio, and a stop share below its crime share. Thus, both the Ratio/Ratio and Share/Share

measures are always less than 1, misleadingly suggesting that the higher-crime group is less intensively policed once you account for crime.[88]

- If there is racial disparity in average stop rates conditional on criminal conduct, disfavoring the higher-crime group, Ratio/Ratio and Share/Share comparisons always misleadingly mask its extent. Specifically, both measures will always be less than a weighted average of the true disproportionalities in the policing of the innocent and of the guilty (D_I and D_G, as defined above).[89] They are also both either less than D_I or less than D_G, and are often less than either one;[90] these relationships are explored further in the next Section. In some cases, they misleadingly reverse the apparent direction of the disparity.[91]

- If policing disparities conditional on criminal conduct disfavor the *lower-crime* group, the Ratio/Ratio measure always exaggerates this disparity,[92] but the Share/Share measure is then more ambiguous. It might actually understate the disparity, especially if the higher-crime group is relatively large. This is because of the "dilution" of the Share/Share measure discussed above, the extent of which depends on relative group size.

- Ratio/Ratio and Share/Share comparisons are more misleading when group crime-rate differences are larger. On the other hand, when there is *no* crime-rate difference, the comparisons are fine.[93]

- Because the problem stems from stops of the innocent, both measures get more misleading when the police are less discerning between guilty and innocent, and when more of the population is innocent. But as the number of stopped innocents moves toward zero, the Ratio/Ratio measure converges on the stop-rate ratio among the guilty (D_G).

None of this means it's wrong to try to "account for crime." But *dividing* by crime rates or shares is a bad way to account for it.

One might wonder whether there is some other justification for using Share/Share or Ratio/Ratio Comparisons—some reason to care whether stop and crime distributions mirror one another, other than the "like cases" principle. Perhaps there is, but none seems obvious, and those who employ Share/Share or

---

[88] See Proof 1 and all the examples in Table A1.

[89] See Proof 3 and Table A2, Cols. 1-4; the notes accompanying Table A2 and Proof 3 discuss several alternative means of weighting the average (the proposition is true for any of them). Proof 3 applies to the Ratio/Ratio measure, but he fact that the Share/Share measure is also lower than any weighted average of D_I and D_G that is greater than 1 follows from a combination of Proofs 2 and 3: if the Ratio/Ratio measure is (while lower) still above 1, the Share/Share measure will be lower yet, and if the Ratio/Ratio measure is below 1, the Share/Share measure will also be below 1.

[90] This follows from the fact that they are lower than the weighted average.

[91] *See* Table A2, Cols. 1-2.

[92] Again, it is always lower (in this case, farther below 1) than D_I and/or D_G, and lower than their weighted average. These points follow from Proof 3.

[93] See the last columns in Tables A1 and A2. Proofs 1, 2, and 3 all assume crime-rate differences.

Ratio/Ratio comparisons have not articulated any. Rather, the comparisons are usually presented as though they speak to the question whether like cases are being treated alike. And audiences are likely to interpret them that way, unless some alternative interpretation is clearly articulated.[94]

I do not think these comparisons are *deliberately* misleading, though they generally mislead in a particular direction: reducing the appearance of overpolicing of higher-crime groups (usually people of color). Both comparison types are routine on both sides of the race-and-policing debate. They seem intuitively correct, and as research on other common mathematical mistakes suggests, it is remarkably easy not to notice when such intuitions are wrong.[95]

D. Estimating Policing Burdens Conditional on Criminal Conduct

How *should* researchers assess racial disparities conditional on criminal conduct, then? If we knew criminal conduct, race, and police interactions for the same sample (including those not stopped), we could directly estimate race gaps in police-interaction rates for any given criminal conduct condition, or overall average disparity conditional on criminal conduct.[96] This may sometimes be possible—in particular, it is possible for surveys that ask respondents to self-report their criminal conduct also to ask about policing outcomes. A few surveys of cohorts of youth have done so, at least with respect to arrest outcomes and certain specific categories of criminal offending. These generally ask respondents about their behavior over

---

[94] If anything, these comparisons suggest an unattractive "group desert" theory—the idea that a group's total policing burden should track its total crime commission, without regard to which specific individuals are guilty. Or perhaps police might argue that it's rational to attach extra suspicion to all members of high-crime groups, and to allocate stops in direct proportion to group crime rates. It's not clear that this approach would be efficient, however, and in the next Part, I argue that it's wrong and unconstitutional for the police to infer criminal propensity based on race. In any case, even if we identified a good reason to allocate stops in proportion to crime shares, it would still be important to recognize that doing so will mean people with the same conduct systematically face different apprehension rates depending on race. Policymakers would have to decide whether to tolerate this breach of the like-cases principle in order to serve some other objective. The literature that employs Share/Share and Ratio/Ratio comparisons has failed to highlight this conflict.

[95] A famous example is the Monty Hall Problem, a brainteaser with a simple but counterintuitive solution that people (even many mathematicians) overwhelmingly resist even after it is explained. *See* John Tierney, *Behind Monty Hall's Doors*, N.Y. TIMES, July 21, 1991 (describing 10,000 letters sent to a magazine columnist protesting her correct explanation, including many from mathematics professors). Humans are so bad at conditional probabilities that (according to a surprisingly large interspecies-comparison literature) we are outperformed on numerous tests by pigeons. *E.g.*, Walter T. Hebranson & Julia Schroeder, *Are Birds Smarter than Mathematicians?*, 124 J. COMP. PSYCHOL. 1 (2013); Edmund Fantino et al., *Teaching Pigeons to Commit Base-Rate Neglect*, 16 PSYCH. SCIENCE 820, 820 (2005).

[96] For example, one could estimate a regression, such as:

$$Stop = \beta_1 * \text{Black} + \beta_2 * \text{Guilty} + \beta_3 * \text{Guilty} * \text{Black} + \alpha$$

Here, β_1 is the additive "black" effect for the innocent; $\beta_1 + \beta_3$ is the additive "black" effect for the guilty; α is the baseline for white innocents, and $\alpha + \beta_2$ is the baseline for white guilty. The predicted probabilities could then be divided to obtain likelihood ratios; for example, $D_I = (\beta_1 + \alpha)/\alpha$. A regression could also include indicators for many possible crime types or degrees of culpability. Removing the interaction terms between "black" and the crime variables would produce an overall average "black" effect conditional on conduct.

some time period (for example, whether they engaged in certain conduct "often," "never," and so forth) and their arrests over the same period.[97] This approach has some limitations, including the general accuracy and sampling concerns about surveys mentioned above, plus likely heightened concerns about statistical power.[98] But well-designed, large surveys at least have the potential to allow individual-level estimation of policing outcome disparities conditional on self-reported conduct.[99]

In the "accounting for crime" literature, however, researchers have typically used police outcome data that comes from official sources, which has advantages, including the fact that police data usually cover the full set of police interactions of interest (e.g., all of a jurisdiction's traffic stops), rather than merely the outcomes for a much smaller surveyed sample. However, crime benchmarks then must come from some other source. Can these kinds of benchmark comparisons be used to assess policing disparities conditional on criminal conduct? And can we translate existing Ratio/Ratio or Share/Share comparisons into the kinds of comparisons we want?

Each of these objectives can be accomplished, but it requires an additional key piece of information not usually included in benchmark-comparison studies: each group's "hit rate," the share of those stopped who are guilty. With this information, we can first translate existing Stop Ratio/Crime Ratio estimates into Stop Ratios for the guilty:

$$D_G = \frac{S_{G1}/P_{G1}}{S_{G2}/P_{G2}} = \frac{S_{G1}/S_1}{S_{G2}/S_2} * \frac{S_1/S_2}{P_{G1}/P_{G2}} = Hit\ Rate\ Ratio * \frac{Stop\ Ratio}{Crime\ Ratio}$$

As this equation shows, D_G always exceeds the Ratio/Ratio measure unless the high-crime group's hit rate is actually below or equal to the low-crime group's. If the crime-rate difference is substantial, this can only happen if the police are *much less accurate* vis-à-vis high-crime group members (Table 3, Col. 4 is an example).

---

[97] *See, e.g.*, David S. Kirk, *The Neighborhood Context of Racial and Ethnic Disparities in Arrest*, 45 DEMOGRAPHY 55 (2008) (analyzing Chicago longitudinal survey of youth cohorts); Robert J. Sampson, *Effect of Socioeconomic Context on Official Reaction to Juvenile Delinquency*, 51 AM. SOC. REV. 876 (1986) (discussing the Seattle Youth Survey). In both these examples, the data were also linked to the study participants' official criminal records to check the accuracy of self-reports of arrests. Such surveys also contain other information about the individual, and studies that use them have generally analyzed a broad set of predictor variables, rather than focusing on crime and race alone.

[98] The sample size needed to estimate differences in arrest rates conditional on a given type of criminal conduct would be larger than the sample size needed to estimate underlying criminal conduct differences, because the outcome is much lower-frequency, given that the great majority of criminal incidents do not result in arrest.

[99] Optimally, to assess disparities among the guilty, surveys should ask the outcome of each reported crime incident. The approach is best suited to crime types for which the respondent is likely to have a clear memory of each incident (not minor, forgettable conduct like loitering). It would be harder to expect respondents to also accurately recount all their *innocent* conduct (for example, how many times they walked down the street *not* carrying a weapon), so similar incident-level analyses would be harder to conduct for overall disparities conditional on criminal conduct or for disparities among the innocent. However, one could model number of arrests for a particular crime type as a function of number of times the respondent self-reports engaging in that crime.

The relationship of D_I, the stop-rate disparity among the innocent, to the Stop Ratio/Crime Ratio measure can also be calculated with the same information:

$$D_I = \frac{S_{I1}/P_{I1}}{S_{I2}/P_{I2}} = \frac{S_{I1}/S_1}{S_{I2}/S_2} * \frac{P_{G1}/P_{G2}}{P_{I1}/P_{I2}} * \frac{S_1/S_2}{P_{G1}/P_{G2}} = \textit{Miss Rate Ratio} * \frac{\textit{Crime Ratio}}{\textit{Innocents Ratio}} * \frac{\textit{Stop Ratio}}{\textit{Crime Ratio}}$$

The "Miss Rate Ratio" is the intergroup ratio of the shares of stops that are *not* "hits," and the Innocents Ratio is the ratio of the groups' innocent populations. Note that the "Crime Ratio/Innocents Ratio" proportion is necessarily over 1 (again, assuming group 1 is higher-crime). The Miss Rate Ratio is usually below 1, barring substantial policing-accuracy differences, but often not dramatically so; if hit rates are below 50%, the Miss Rate Ratio is less disproportionate than the Hit Rate Ratio.[100] But it always cuts in the opposite direction. Hence, if D_G is *lower* than the Stop Ratio/Crime Ratio measure, D_I must be higher (because both multipliers in the equation above would be above 1). Simplifying the equation:

$$D_I = \textit{Miss Rate Ratio} * \frac{\textit{Stop Ratio}}{\textit{Innocents Ratio}}$$

The Stop Ratio and Innocents Ratio could equivalently both be replaced with ratios of stop *rates* and innocence *rates* (the population terms cancel).

Alternatively, we can calculate D_G and D_I if we know the Stop Shares and Crime Shares plus the hit rates for each of two groups. The Stop Ratio and Crime Ratio can first be calculated from the shares,[101] and then the formulas above can be applied.

As it happens, race-specific hit rates are frequently reported in the policing-disparity literature. Some commentators treat racial equality in hit rates as evidence that policing disparities have no crime justification, effectively using hit rates as a proxy for crime rates.[102] But in reality, "reported hit rates typically exceed the range of plausible crime rates,"[103] and the relationship between hit rates and crime rates could differ across races. For example, the Table A2 examples all assume a Crime Ratio of 2, but the Hit Rate Ratios vary from 0.56 to 2.98.

A recently prominent line of research interprets equal hit rates in just the opposite way, to imply *lack* of police racial bias. This approach has complicated and problematic assumptions, which are examined in detail in Part III. For now, note that equal hit rates are compatible with stark racial disparities in policing rates conditional on criminal conduct.[104] Conversely, unequal hit rates are the inevitable

[100] For example, suppose the black and white hit rates are 0.2 and 0.1 respectively (hit-rate ratio = 2). Then the miss rate ratio is 0.8/0.9, or 0.89. Low hit rates are common for investigative stops; the NYPD claims a hit rate of 12% for its stop-and-frisk policy (but less than 1% involve weapons).

[101] Dividing any group's crime share (or stop share) by any other's gives a ratio of crimes (or stops).

[102] *E.g.*, DAVID A. HARRIS, PROFILES IN INJUSTICE 79-84 (2002).

[103] R. Richard Banks, *Beyond Profiling: Race, Policing, and the Drug War*, 56 STAN. L. REV. 571 (2003).

[104] The test's leading advocate, Nicola Persico, has long acknowledged that the predicted equilibrium under "unbiased" policing could, depending on various conditions, be highly "unfair." Persico, *Racial Profiling, Fairness, and Effectiveness of Policing*, 92 AM. ECON. REV. 1472, 1479 (2002).

result of *equal* policing conditional on criminal conduct, if crime rates differ. These points are illustrated by the Hit Rate Ratios in the Appendix tables.[105]

But if we have hit-rate *and* crime-rate information that we trust, we can use that information in a different way, to estimate the quantities of interest D_G and D_I. This is a big "if." Even setting aside the crime-benchmarking challenges discussed above, accurate hit rates may be elusive. Hit-rate studies often use arrests to measure hits, but this is highly problematic. Police have broad arrest discretion, and it is nonsensical to assume, when assessing racial disparities in stops or searches, that there is no unjustified racial disparity in arrests. Nor are convictions a good "hit" measure—conviction depends on prosecutorial discretion and on police testimony. Police records of whether contraband was found could likewise be affected by disparities in search intensity or decisions to look the other way.

This problem also bedevils all existing uses of "hit rates" to assess racial disparities in policing, however.[106] If researchers *do* trust their hit measure, it might as well be employed in a theoretically sensible way. And some hit measures may be reasonably reliable because police have little discretion—for example, Breathalyzers that automatically record results, or crimes that are so serious that police almost never fail to arrest. How much discretion police have will not be obvious from the data; researchers need a strong qualitative sense of how policing works in the context they are studying. When researchers doubt hit measures' accuracy, they can use assumptions about a plausible range of hit rates to estimate bounds on D_G and D_I.

David Thacher, in an unpublished 2002 paper, came closest to this approach. Analyzing NYPD's stop-and-frisk policy, for each racial group, he counts all stops not resulting in arrests as stops of innocents, and divides that number by the total group population. He finds that the probabilities of an innocent person being stopped are 0.6% for whites, 4.2% for blacks, and 2.9% for Hispanics.[107] In addition to assuming that arrest accurately gauges innocence, Thacher apparently assumes guilt rates are so negligible that the total population can substitute for the innocent population. Both assumptions are likely conservative,[108] yet the disparities are

[105] As the Table A1 examples illustrate, if the same shares of the guilty and innocent are stopped from each group, the group with a higher guilt rate overall must have a higher guilt rate among stops. Meanwhile, in Table A2, black hit rates are higher in almost every hypothetical (which the advocates of the "outcome test" would interpret to suggest irrational bias *favoring* black suspects), including in Columns 1-3, in which black pedestrians are actually much more heavily policed conditional on criminal conduct. Moreover, the hit rate ratio in Column 1 is very similar to the one in Column 5, even though the directions of the policing disparities are reversed. A lower black hit rate is obtained only when we assume policing that is *far* less discerning for black pedestrians (Col. 4).

[106] Decio Coviello & Nicola Persico, *An Economic Analysis of Black-White Disparities in NYPD's Stop-and-Frisk Program*, Working Paper (2013), 13-14, acknowledge this problem, but argue that there is no evidence police use arrest discretion in racially disparate ways, because arrest rates are uncorrelated by race after controlling for "crime type" recorded on post-stop forms. This does not help. What the officer writes down is discretionary, and could be a post hoc rationalization.

[107] Thacher, *supra* note 35, at 37 tbl. 1.

[108] The use of arrest to measure hits is conservative if arrest disparities run in the same direction as stop disparities; the use of the whole population to substitute for innocents is conservative assuming (as NYPD has contended) that black New Yorkers have a higher crime rate.

dramatic. The stop-and-frisk policy doesn't look good in terms of disparity in burdens on the innocent.

But what about disparities affecting the guilty? Thacher does not address this question for stop-and-frisk, perhaps because doing so would require crime benchmarks. One cannot simply assume away the existence of the guilty population when evaluating their stop probabilities. Instead, Thacher praises Lamberth's Turnpike study for having "properly estimated the distribution of burdens on the guilty."[109] Is this praise merited? Well, recall that Lamberth found that virtually all drivers were guilty of (minimal) speeding; if minimal speeding is the proper measure of "guilt" (which has been sharply contested), then there would be very few stops of the innocent, so the problems outlined in the previous section wouldn't emerge.[110] But Thacher does not explain how disparities affecting the guilty should be estimated when guilt is less universal, nor how disparities affecting the innocent should be estimated when guilt rates are nontrivial.

Both, however, can be estimated, albeit not easily. When we lack either a good crime benchmark or a good hit measure, the best we may be able to do is to offer a reasonable range of estimates based on assumptions. But even if the questions posed by the "like cases" framework are empirically challenging, they are the right questions if we want to "account for crime" in a normatively meaningful way.

III. Measuring Racial Discrimination in Police Decision-Making

What if we want to go beyond asking whether policing disparities are explained by crime commission, and instead ask whether they are driven by racial discrimination? This Part assesses this question, beginning by asking why it is constitutionally central and morally important. Section B examines causal inference problems that complicate the task. Section C considers the literature's main approaches: the "hit-rate" test, regression studies of stops and post-stop outcomes, studies exploiting variations in ability to observe race, and lab studies of implicit biases. Finally, Section D argues for a new strategy—the use of "testers," like those used in other antidiscrimination enforcement and research contexts—and addresses how to address practical, safety, and ethical concerns specific to the policing context.

A. The Constitutional and Moral Case Against Racial Profiling

While many causes of policing disparity raise policy concerns, the role of governmental discrimination is the key question posed by equal protection doctrine. Current doctrine precludes constitutional challenges solely premised on racially disparate impact[111] or discrimination by private actors like witnesses.[112] But, as I show

---

[109] *Id.* at 24.

[110] Lamberth actually divided each group's stop share by its share of all drivers on the road. If almost everyone is guilty, this is close to a Stop Share/Crime Share comparison.

[111] Washington v. Davis, 426 U.S. 229, 239-42 (1976).

[112] One could argue that when the police give effect to such private discrimination by carrying out stops and arrests, state action is generated. *Cf. Shelley v. Kraemer*, 334 U.S. 1 (1948) (barring judicial enforcement of private racially restrictive covenants). But doctrinally, this is likely a nonstarter. *See* Don Herzog, *The* Kerr *Principle*, 105 MICH. L. REV. 1, 40 (2006) (dismissing a similar hypothetical

here, police racial discrimination essentially always violates the Equal Protection Clause. It is, however, difficult to prove, which makes effective empirical strategies especially important—for parties to litigation, for courts, for Congress should it legislate under the Fourteenth Amendment, and for police departments who wish to comply with the Constitution and avoid litigation. Moreover, the disparate treatment question matters for other normative reasons as well.

Although there is a strong scholarly consensus that racial profiling *should* be considered unconstitutional, scholars often question whether the Supreme Court agrees. Many have critiqued the Court for leaving the door open to police reliance on race.[113] These critics have grounds for frustration: the Court has avoided squarely deciding the Fourteenth Amendment issue and has meanwhile foreclosed Fourth Amendment strategies. Moreover, lower courts have been unwilling to second-guess police reliance on race-specific suspect identifications, even in extreme cases.[114] Still, broader equal protection doctrine leaves little ambiguity. Racial profiling (by which I mean reliance on conscious or subconscious racial generalizations about criminality, as opposed to specific suspect identifications) clearly violates the Equal Protection Clause as the Court has consistently interpreted it.[115]

Scholars examining the relevant constitutional doctrine have mainly focused on the Court's numerous adverse Fourth Amendment precedents.[116] These include *Whren v. United States*, holding that a traffic stop provides probable cause for a vehicle search even if the traffic violation was a mere pretext,[117] and *United States v. Brignoni-Ponce*, which suggested that Mexican appearance might provide reasonable suspicion

---

extension of *Shelley*); David Strauss, *Discriminatory Intent and the Taming of* Brown*,* 56 U. CHI. L. REV. 935, 966-83 (1989) (reviewing post-*Shelley* doctrine). It would also set a difficult standard for police, who may not know when witnesses are racially biased.

[113] *E.g.*, Johnson, *supra* note 5, at 1006; Delores Jones-Brown & Brian A. Maule, *Racially Biased Policing*, in RACE, ETHNICITY, AND POLICING, *supra* note 6, at 141-43; Paulaumus et al., *supra* note 43, at 242-43; Evan Gerstman & Christopher Shortell, *The Many Faces of Strict Scrutiny*, 72 U. PITT. L. REV. 1, 46-50 (2001); Alschuler, *supra* note 27, at 164-66; Angela J. Davis, *Race, Cops, and Traffic Stops*, 51 U. MIAMI L. REV. 425, 442-43 (1997).

[114] Notoriously, in *Brown et al. v. City of Oneonta*, 221 F.3d 329 (2000), the Second Circuit upheld the interrogation of 200 black men based on a white victim's description of a black male assailant. For critiques, see Gerstman & Shortell, *supra* note 113, at 47; Alschuler, *supra* note 27, at 179-92. *Oneonta*'s facts were shocking; one can hardly imagine a race-reversed scenario in which 200 white men were stopped. But plaintiffs do not win equal protection cases based on what one can and can't imagine, even if sometimes they should. *See United States v. Armstrong*, 517 U.S. 456 (1996) (requiring an actual comparison group). Broad sweeps like this may still be attacked for lack of Fourth Amendment individualized suspicion; the court in *Oneonta* allowed that argument to proceed. 221 F.3d at 334. In any event, courts consistently distinguish racially specific suspect descriptions from behavioral generalizations about groups. *See* R. Richard Banks, *Race-Based Suspect Selection and Colorblind Equal Protection Doctrine and Discourse*, 48 UCLA L. REV. 1075, 1078-80 (2001) (critiquing this distinction).

[115] The Sixth Circuit has gotten this issue wrong, however. *E.g.*, United States v. Travis, 62 F.3d 170, 174 (1995); *see* Alschuler, *supra* note 27, at 178-79 (critiquing these cases).

[116] *E.g.*, Jones-Brown & Maule, *supra* note 113, at 140-57; Jeffrey A. Fagan et al., *Street Stops and Broken Windows Revisited*, in RACE, ETHNICITY, AND POLICING, *supra* note 6, at 312-13; Johnson, *supra* note 5, at 1006-08.

[117] 517 U.S. 806, 813-16 (1996).

of an immigration violation when combined with other factors, but not alone.[118] But these cases do not implicate equal protection claims. *Brignoni-Ponce* sent a confusing signal (why suggest that ethnicity may be relevant to Fourth Amendment analysis if its consideration is barred by the Fourteenth?), but it does not trump the many equal protection cases striking down decision-makers' use of race even alongside other factors.[119] Many scholars have critiqued the doctrinal separation of Fourth and Fourteenth Amendment objections to racial profiling,[120] but its upside is that adverse holdings on the former do not preclude the latter.[121]

The Supreme Court has never decided whether racial profiling violates the Fourteenth Amendment, but to say no, it would have to upend decades of doctrine. The key line of cases concerns the prohibition on "statistical discrimination." The Supreme Court has consistently held that otherwise-impermissible discrimination cannot be justified based on group generalizations, even if those generalizations are empirically accurate. Instead, individuals must be treated as individuals.[122]

For example, in *Craig v. Boren*, the Court struck down a law applying different drinking ages to men and women. It was unmoved by studies showing that young men drove drunk at ten times the rate of young women, because these findings lumped all young men together.[123] Similarly, the Court has struck down governmental reliance on gendered or racial generalizations about learning styles,[124] juror voting,[125] and workforce participation.[126] All these generalizations had statistical support, but the Court made clear that this doesn't matter: basing disparate treatment on groups' typical tendencies is unfair to atypical individuals. The Court has carved out exceptions only for physical sex differences relating to pregnancy.[127] It has never made exceptions for generalizations about behavior, and it would be shocking if it did so for racial generalizations about criminal tendencies.

---

[118] 422 U.S. 873, 885-87 (1975); *see* United States v. Martinez-Fuerte, 428 U.S. 543 (1976).

[119] *E.g.,* Arlington Heights v. Metropolitan Housing Dev. Corp., 429 U.S. 252, 265-66 (1977); *see* Gross & Barnes, *supra* note 11, at 740.

[120] *E.g.*, Alschuler, *supra* note 27, at 193 (reviewing commentary); David A. Sklansky, *Traffic Stops, Minority Motorists, and the Future of the Fourth Amendment*, 1997 SUP. CT. REV. 271, 309-29 (1997).

[121] Scholars have suggested that *Whren* and related cases green-light racial profiling in car searches. *See* Jones-Brown & Maule, *supra*, at 153-57 (also citing *Maryland v. Wilson*, 519 U.S. 408 (1997) and *Atwater v. City of Lago Vista*, 532 U.S. 318 (2001)); Paulhaumus et al., *supra* note 43, at 242-43; Fagan et al., *supra* note 116, at 312. This is likely often true in practice, because allowing pretextual justifications makes it harder to prove racial discrimination. But *Whren* did not suggest it would be *legal* to rely on race—it suggested otherwise. 517 U.S. at 813 ("[T]he constitutional basis for objecting to intentionally discriminatory application of laws is the Equal Protection Clause, not the Fourth Amendment").

[122] *See* Sonja B. Starr, *Evidence-Based Sentencing and the Scientific Rationalization of Discrimination*, 66 STAN. L. REV. 803, 823-29 (2014) (analyzing these cases).

[123] 429 U.S. 190, 202-04 (1976).

[124] United States v. Virginia, 518 U.S. 515, 532-34 (1996).

[125] J.E.B. v. Alabama ex rel. T.B., 511 U.S. 127, 130-31 (1994); Batson v. Kentucky, 476 U.S. 79, 90, 97-98 (1986).

[126] *See* Weinberger v. Wiesenfeld, 420 U.S. 636, 645 (1975); Frontiero v. Richardson, 411 U.S. 677, 690-91 (1973).

[127] *See* Tuan Anh Nguyen v. I.N.S., 533 U.S. 53, 68 (2001).

This line of cases should be fatal to any attempt to justify racial profiling by arguing that profiles are empirically supported. This is so even assuming crime prevention is a compelling state interest.[128] In none of the cases reviewed above did the Court assess whether the statistical generalization in question established an important government interest. Rather, the prohibition on statistical discrimination is best understood to constrain the kinds of reasoning that the government can offer to establish its interest. Otherwise, the law in *Boren* might well have survived scrutiny, for example. The government clearly has an important interest in preventing drunk driving—yet it was barred from using statistical evidence to show a relationship between that interest and the gender classification.

Moreover, law enforcement bodies have generally agreed that profiling is illegal. For example, in 2003 DOJ declared it "absolutely prohibited."[129] The remaining ambiguity in the case law may therefore be irrelevant in practice. Modern police departments don't defend racial profiling. They deny that they engage in it.[130] Most profiling lawsuits have settled on terms prohibiting it.[131]

But if the Fourteenth Amendment argument is doctrinally well supported and not in practice contested, why has the Fourth Amendment played a more prominent role in profiling litigation? The key problem is evidentiary: it's hard for litigants to prove racial profiling, and especially to prove that it affected any specific case.[132] Individual criminal defendants raising selective-enforcement defenses face very steep hurdles.[133] In federal criminal cases, even getting discovery on the issue is notoriously difficult.[134] And even if defendants can show a pattern of discrimination, they must

---

[128] *See* United States v. Virginia, 518 U.S. 515, 531 (1996) (listing prohibition on gender generalizations and a substantial relationship to important government interests as separate requirements).

[129] U.S. DEP'T OF JUSTICE, FACT SHEET: RACIAL PROFILING 3 (2003), http://www.justice.gov/archive/opa/pr/2003/June/racial_profiling_fact_sheet.pdf.

[130] *E.g.*, *Melendres v. Arpaio*, 695 F.3d 990, 995 (9th Cir. 2012) (describing Arizona sheriff's defense to equal protection suit: "Defendants do not engage in racial profiling"); News Hour with Jim Lehrer, Aug. 13, 2013 (quoting NYPD Commissioner: "We do not engage in racial profiling. It is prohibited by law [and] by our own regulations."); Sho Wills, *Chicago, New York Officers Spar Over Stop-and-Frisk Policy*, CNN, http://www.cnn.com/2013/08/14/us/new-york-chicago-stop-frisk/ (Aug. 14, 2014) (quoting Chicago PD spokesman); Greg Risling, *DOJ Finds Two LA Sheriff's Stations Discriminating*, ASSOC. PRESS (June 28, 2013) (describing L.A. County Sheriffs' response to DOJ investigation); *All Police to Attend Racial Profiling Class*, BURLINGTON CNTY. TIMES (June 29, 2005) (stating that all New Jersey police officers must attend training teaching unconstitutionality of profiling); Jane Prendergast, *Officers' Hearts Hold Racial Profiling Solution, Chief Says*, CIN. ENQUIRER (Mar. 6, 2011) (quoting Cincinnati police chief: Profiling "is not only wrong, it's unconstitutional. It's illegal. We know that. We teach that."); Letter from S.C. Kitchen to Assistant U.S. Attorney General Thomas E. Perez, (September 26, 2013), http://www.timesnewshosting.com/docs/johnson.pdf (denying profiling).

[131] *See* Gross & Barnes, *supra* note 11, at 741-43.

[132] *See id.* at 653-57, 741; David Rudovsky, *Law Enforcement by Stereotypes and Serendipity*, 3 U. PA. J. CONST. L. 296, 322-29 (2001); Johnson, *supra* note 5, at 1063-64.

[133] *See, e.g.*, KENNEDY, *supra* note 30, at 354 ("Research has uncovered no cases" of convictions overturned for selective prosecution, as of 1997).

[134] The Supreme Court has required "some evidence" of "differential treatment of similarly situated members of other races." *Armstrong*, 517 U.S. at 465-67. Identifying a "similarly situated" group is notoriously difficult, and may be more so in policing cases: police keep no "records of instances in which they could have stopped a motorist…but did not." Davis, *supra* note 113, at 437-38.

also show that it affected their cases specifically. Statistical evidence about broader patterns almost never clears this hurdle alone, though it might help in combination with case-specific qualitative evidence.[135] For these reasons, the best prospects for Fourteenth Amendment challenges to succeed are in civil cases (class actions or government enforcement actions), in which the pattern of discrimination is the issue. Such cases turn centrally on statistical evidence.

Some commentators, while acknowledging the legal importance of the disparate-treatment question, have dismissed its moral importance. Thacher, for example, describes the focus on "racial profiling" as a parochial concern of lawyers—a distraction from "substantive" equality.[136] In fact, however, this call for a move beyond colorblindness echoes a view that has long been common in *legal* scholarship beyond the policing context: that equality law should primarily target group subordination, not forbidden classifications.[137]

I generally sympathize with this antisubordination view. But whether the police racially discriminate is not "merely" a legalistic concern. Racially disparate treatment adds a substantively meaningful dimension of harm. Critics of racial disparities in policing (not just lawyers) have emphasized the role of discrimination, "racial profiling," or just "racism."[138] This framing has cultural resonance and moral force.[139] For the government to generalize that people of color are dangerous, and to specifically target them for surveillance and arrest, is expressively and morally noxious, especially because such generalizations have a painful history in our culture.[140] I am not suggesting that one should not acknowledge racial differences in crime rates. But the poisonous aspect is using those differences to justify ignoring differences *within* groups, making law-abiding citizens "pay for fears generated by criminals with which they are lumped by dint of color."[141]

---

[135] In principle, strong statistical evidence could allow an inference that the defendant *probably* would not have been stopped but for race. But courts have resisted this sort of reasoning, *see* Harcourt, *supra* note 13, at 1278, just as they are often uncomfortable inferring individual causation from statistics in other kinds of cases, *see* Laurence H. Tribe, *Trial by Mathematics: Precision and Ritual in the Legal Process*, 84 HARV. L. REV. 1329, 1349-51 (1971). In *McCleskey v. Kemp*, 481 U.S. 279 (1987), the Court refused to allow a defendant's challenge to his capital sentence to rest solely on statistical findings of racial disparity in death penalty administration. This holding emphasized deference to prosecutors and juries, and could possibly be distinguished in a challenge to police racial profiling; the Court has upheld equal protection claims based on statistical evidence in some other criminal-law contexts. *E.g.*, Castaneda v. Partida, 430 U.S. 482 (1977); Whitus v. Georgia, 385 U.S. 545 (1967).

[136] Thacher, *supra* note 35, at 26; *see supra* note 40 (addressing other critics of discrimination focus).

[137] *E.g.*, Jack M. Balkin & Reva B. Siegel, *The American Civil Rights Tradition: Anticlassification or Antisubordination?*, 58 U. MIAMI L. REV. 9 (2003).

[138] *E.g.*, *The Targeting of Young Blacks by Law Enforcement: Ben Jealous in Conversation with Jamelle Bouie*, AM. PROSPECT, Fall 2014; Rockeymoore et al, *supra* note 3*;* Statement of Rep. John Lewis, http://johnlewis.house.gov/press-release/rep-john-lewis-shooting-michael-brown-and-events-ferguson-missouri.

[139] *See* BLANK ET AL., *supra* note 84, at 103 (stating more generally that "the broader public vision of what discrimination means [is] the treatment of two (nearly) identical people differently").

[140] *See, e.g.*, KENNEDY, *supra* note 30, at 16; Patricia Williams, *Spirit-Murdering the Messenger*, 42 U. MIAMI L. REV. 127, 129-30 (1987); Barack Obama, Remarks on Trayvon Martin, July 19, 2013.

[141] KENNEDY, *supra* note 30, at xi.

The harms of racially disparate policing are thus often substantially connected to racial targeting, not just to police interactions *per se*. Perceptions of police racism also deeply undercut trust in the police in black communities, which may undermine police effectiveness.[142] In short, racial discrimination may be just one morally troubling cause of racial disparity, but it's an important one.

B. Race and the Problem of Causal Inference

Identifying the effects of racial discrimination is even harder than assessing racial disparities in policing of people with like criminal conduct. The latter inquiry need not involve causal claims; the racial discrimination question does.[143] Answering it requires researchers to disentangle not only the role of crime but also all other potentially confounding variables. This is challenging, because race is not a "treatment" subject to manipulation. Its effects are intertwined with each individual's other attributes and life experiences—which may themselves have been influenced by race. Scholars have therefore debated whether the language of causal inference can be meaningfully applied to race at all.[144] Perhaps we can't sensibly ask how a person's outcome would have differed but for her race if her entire life would have been different in that counterfactual.

This conceptual hurdle is not insuperable, however. Usually, when we ask causal questions about race ("Do police officers stop more African-Americans because of their race?"), we're not asking about race's total, lifelong effects, but about racial discrimination in a particular decision process.[145] The counterfactual is how the decision-maker would have responded if she had instead encountered another person of a different race but otherwise similar relevant characteristics. James Greiner and Donald Rubin have suggested referring to "perceived race" to highlight this point.[146] In my view, while the point is sound, there's little harm in the shorthand "effect of race," provided we are clear on what it means.[147]

But isolating "effects of race" even in this narrower sense is difficult. It is not just that race resists experimental manipulation; for any given individual, it does not vary naturally either. Immutable traits defy observational researchers' best tools for causal inference, such as panel designs (which follow individuals before and after a treatment) or quasi-experiments exploiting shocks to a treatment.[148] Instead,

---

[142] *See id.* at 151-53; Tom R. Tyler & Jeffrey Fagan, *Legitimacy and Cooperation*, in RACE, ETHNICITY, AND POLICING, *supra* note 6, at 102-04; Weitzer & Tuch, *supra* note 2, at 1017-18.

[143] *See, e.g.*, Lincoln Quillian, *New Approaches to Understanding Racial Prejudice and Discrimination*, 32 ANN. REV. SOCIOLOGY 299, 302 (defining "discrimination" as "the causal effect of race").

[144] D. James Greiner & Donald B. Rubin, *Causal Effects of Perceived Immutable Characteristics*, 93 REV. ECON. & STAT. 775, 783-84 (2011); Maya Sen & Omar Wasow, *Reconciling Race and Causation* (2014), https://www.princeton.edu/csdp/events/Wasow04092014/race_causation_4.pdf.

[145] If one *did* want to examine "the effect of dynamic, cumulative discrimination," BLANK ET AL., *supra* note 84, at 226, the strategies discussed here wouldn't much help; raw disparity estimates might.

[146] Greiner & Rubin, *supra* note 144, at 775.

[147] "Effect of perceived race" is itself a shorthand; every individual has been affected by perceptions of her race her whole life.

[148] Quasi-experimental designs *can* be used to assess changes or differences in racial disparity. *See, e.g.*, Sonja B. Starr & M. Marit Rehavi, *Mandatory Sentencing and Racial Disparity*, 123 YALE L.J. 2 (2013)

researchers must use methods—such as regression, reweighting, and matching—that share a core limitation: their ability to support causal inferences depends on the ability to observe the potentially confounding variables.[149] Because one can never really exclude the possibility of omitted variables, careful researchers often refer to the race gaps that remain after controlling for observed variables simply as "unexplained," rather than claiming proof of discrimination.[150]

Still, neither policy analysis nor law requires definitive answers. For policy purposes, strong causal identification would be great, but even an analysis with weaker identification can usefully narrow down the possibilities; theoretically informed discussion can then guide interpretation of unexplained gaps. In civil litigation, the traditional burden of proof requires that the factfinder believe the best interpretation of the evidence is that discrimination *probably* had some effect. After all, non-statistical evidence of causation (and other contested facts) is often also open to multiple interpretations. While courts have sometimes demanded more clarity out of statistical evidence, certainty or even near-certainty is too much to ask for.[151] Moreover (though there is no clear doctrine on this), it should be unnecessary to rule out every theoretically possible confounding variable in order to support an equal protection claim. Instead, the key question should be whether the explanations for disparities that the police department gives hold up.[152]

In the remainder of this discussion, I assume that strong causal identification is the ideal goal of research on police racial discrimination. However, I also examine what we can learn from observational research that falls somewhat short of this goal.

C. Current Methods of Measuring Police Racial Discrimination

How might one isolate the effects of police racial discrimination? In this Section, I consider several approaches from the literature: the "hit-rate" test, neighborhood-level studies of initial stops, studies of post-stop decisions, studies exploiting variations in decision-makers' information about race, and lab experiments on

(using regression discontinuity design); David S. Abrams et al., *Do Judges Vary in their Treatment of Race?*, 41 J. LEGAL STUDIES 347 (2012) (exploiting random assignment of judges).

[149] Regression models express some functional relationship between variables; the function is fit to data to solve for its parameters, identifying each covariate's association with the outcome when the others are held constant. Reweighting and matching are methods of rendering groups comparable in characteristics before comparing their outcomes.

[150] *See, e.g.*, Quillian, *supra* note 143, at 303.

[151] The case that set the hardest standard was *McCleskey*, in which the Supreme Court, invoking deference to prosecutors and jurors, held that "exceptionally clear proof" of discrimination was required to support a challenge to the death penalty. But it is not obvious that *McCleskey*'s reasoning applies to policing at all, *see supra* note 135, or that it applies to civil lawsuits alleging a pattern of discrimination. *McCleskey* (and every federal appellate case following it) centers on the problem of inferring discrimination in an individual criminal case from a broader statistical pattern.

[152] In petit and grand jury discrimination cases, the Supreme Court has required the state to articulate reasons for its decisions and "stand or fall on the plausibility" of those reasons. Miller-El v. Dretke, 545 U.S. 231, 252 (2005); *see* Casteneda v. Partida, 430 U.S. 482, 494-95 (1970) (holding that this burden-shifting can be triggered by statistical evidence of disparate *impact*); *see also McDonnell-Douglas Corp. v. Green*, 411 U.S. 792 (1973) (establishing similar requirement in Title VII cases).

implicit biases. While I do not return to the crime-benchmark comparisons addressed in Part II, such comparisons have also often been presented as methods for assessing "racial profiling"—implying that if crime differences do *not* explain policing differences, racial discrimination does.[153] This inference could be reasonable, but only absent plausible alternative explanations.[154]

1. The Hit-Rate Test

In recent years, the most prominent strategy for causal inference about race and policing has been the hit-rate test.[155] In Part II.D, I showed briefly that this test does not address equality in policing conditional on criminal conduct. Here, I show that it also fails to identify racial discrimination—or anything we likely want to know.

The method posits that unbiased police seek to maximize the probability of detecting crime, and thus shift their attention toward groups with higher "hit rates."[156] Those groups then reduce their crime rates, which leads to police stopping them less, and so forth. At equilibrium, all groups have the same hit rates (and crime rates). On this view, unequal hit rates suggest racial bias: police who fail to shift stops to the higher-rate group must irrationally prefer to stop the other group.[157] Equal hit rates suggest no bias.[158] The underlying theory distinguishes "taste-based" discrimination (prejudice) from statistical discrimination (use of race as a proxy for a legitimate consideration). Economists sometimes defend the latter as efficient.[159]

But even if it is, it wouldn't render racial discrimination constitutional. As the previous Section outlined, the Supreme Court has consistently rejected the distinction between taste-based and statistical discrimination: it is no defense that police were "correct" to consider race. And the model's vision of "unbiased" policing does not merely *allow* unconstitutional statistical discrimination—it *requires* it. Unbiased police are expected to track hit rates by race and to target high-rate racial groups until rates equalize. Failure to do so is interpreted as irrational bias against the

---

153 For example, the various Turnpike studies were designed to test racial profiling allegations.

154 For instance, perhaps highway troopers are not able to observe much besides speed and race.

155 *See* Engel, supra note 40, at 16 (describing this test as "the analytical strategy of choice [for] many researchers, police administrators, court officials, citizen groups, and other stakeholders").

156 The seminal paper is John Knowles et al., *Racial Bias in Motor Vehicle Searches: Theory and Evidence*, 109 J. POL. ECON. 203 (2001); *see also* Nicola Persico & Petra Todd, *Generalizing the Hit Rates Test for Racial Bias in Law Enforcement, with an Application to Vehicle Searches in Wichita*, 116 ECON. J. F351 (2006); Ruben Hernández-Murillo & John Knowles, *Racial Profiling or Racist Policing? Bounds Tests in Aggregate Data*, 45 INT'L ECON. REV. 959 (2004); Persico, *supra* note 104.

157 *E.g.*, Hernández-Murillo & Knowles, *supra* note 156, at 959; Sean Childers, Note, *Discrimination During Traffic Stops*, 87 N.Y.U. L. REV. 1025 , 1041 (2012).

158 *E.g.*, Persico & Todd, *supra*, at F361; Kate Antonovics & Brian G. Knight, *A New Look at Racial Profiling*, 91 REV. ECON. & STAT. 163, 171-72 (2009); Joseph A. Schafer et al., *Decision-making in Traffic Stop Encounters*, 9 POLICE Q. 184, 200 (2006).

159 *See* Edmund S. Phelps, *The Statistical Theory of Racism and Sexism*, 62 AMERICAN ECONOMIC REV. 659 (1972). Economists have debated the conditions under which statistical discrimination is efficient. *E.g.*, Stewart Schwab, *Is Statistical Discrimination Efficient?*, 76 AM. ECON. REV. 228 (1986); Peter Norman, *Statistical Discrimination and Efficiency*, 70 REV. ECON. STUD. 615 (2003).

lower-rate group. It would thus be wholly inappropriate for courts to use the approach to test for unconstitutional discrimination, as some propose.[160]

The approach is also unsatisfying on policy grounds: it in fact tests neither equality nor efficiency. As to equality, the test embraces race-based policing and toxic racial generalizations about criminality, and as discussed in Part II, equal hit rates are completely consistent with stark differences in treatment of those with the same criminal conduct. As to efficiency, the test does not speak to whether police are minimizing crime, much less maximizing total social utility. Rather, under strong assumptions, it measures whether police maximize arrests. Maximizing arrests and minimizing crime are not synonymous. A race-conscious crime-minimization strategy would have to focus not on hit rates, but on *responsiveness* to policing. As Bernard Harcourt has shown, policing that targets higher-crime groups can *increase* net crime if those groups respond less favorably to police presence.[161]

Indeed, some of the test's leading proponents have acknowledged this point, but have argued that officers' career incentives favor maximizing arrests, not reducing crime, so from their perspective, maximizing hit rates is rational.[162] Descriptively, this could be right. But why should researchers or policymakers share this objective? Essentially, the hit-rate model tests whether officers are racially discriminating in a way that serves their career goals, and not in other ways that do not. A police department that "passes" that test should hardly be proud of it.

Even as a test of whether police maximize arrests, the approach may tell us little, because it relies on extremely strong and dubious assumptions. First, it assumes the hit measure is itself untainted by discrimination, a problem discussed in Part II. Second, it assumes police accurately track hit rates by race and shift stop patterns accordingly. This is probably rarely true.[163] Third, the model assumes that potential criminals know their stop probabilities, update that information when the police shift stop patterns, and change their behavior accordingly. But potential criminals actually face highly ambiguous information about detection rates,[164] and when those rates change, they often don't adjust their behavior proportionally.[165] Fourth, the model also assumes officers simultaneously track hit-rate differences across all *other* characteristics and behaviors that they observe and adjust stop rates accordingly, and

---

[160] Floyd v. City of New York, 813 F.Supp.2d 417, 450-53 (S.D.N.Y. 2011) (describing NYPD's argument); Nicola Persico & David A. Castleman, *Detecting Bias: Using Statistical Evidence to Establish Intentional Discrimination in Racial Profiling Cases*, 2005 U. CHI. LEGAL F. 217, 233 (2005). Others have correctly responded that taste-based discrimination is only one subset of disparate treatment. *E.g.*, Ian Ayres, *Three Tests for Measuring Unjustified Disparate Impacts in Organ Transplantation*, 48 PERSP. IN BIOL. & MED. S68, S80 (2005). Engel, *supra* note 40, at 3, puzzlingly equates statistical discrimination to disparate *impact* discrimination; it is disparate treatment.

[161] Harcourt, *supra* note 13, at 1296-1307.

[162] Persico, *supra* note 104, at 1473-74; Coviello & Persico, *supra* note 106, at 6-11 (explaining that the test doesn't work for allocation of police among precincts, which likely *does* aim to reduce crime).

[163] Only recently have some departments tracked stops and searches by race, and it's unlikely that officers regularly check these figures and understand how to interpret them.

[164] *See* Thomas A. Loughran, *On Ambiguity in Risk Perceptions*, 49 CRIMINOLOGY 1029 (2011); Lance Lochner, *Individual Perceptions of the Criminal Justice System*, NBER Working Paper 9474, at 1 (2006).

[165] Lochner, *supra* note 164, at 29; *see also* Engel, *supra* note 40, at 25 (raising a similar criticism).

that individuals respond such that equal-hit-rate equilibria are produced across all those characteristics and behaviors. This assumption is crucial to solving the "inframarginality problem" that would otherwise invalidate the model.[166]

In addition, the model does not overcome the key problem of omitted variable bias. The discrimination being measured could be a "taste" for an unobserved race-correlated trait, not race itself.[167] Finally, the model predicts that each group's hit rate will equal its crime rate,[168] but this is demonstrably false: observed hit rates often *vastly* exceed plausible underlying crime rates.[169]

In short, the hit-rate approach has little but mathematical elegance to recommend it. Several of the concerns raised here have been raised by others and conceded by the model's creators—but these concessions are quite serious, and go to the heart of whether we should trust the model. The test's continued prevalence is puzzling, and presumably stems from the perceived absence of viable alternatives.[170] But in my view, several of the approaches reviewed below are superior despite their limitations; and in Section D, I propose a new alternative.

2. Neighborhood-Level Studies of Stop Probability

Again, for the most part we lack individual-level, general-population data about underlying criminal conduct, a problem for studies of initial stop probability. It might be possible to construct data sources (especially self-report surveys) that include crime information, stop information, and other potential confounding variables; the self-report surveys discussed in the previous Part have focused on arrests. For now, however, regression studies of initial stop probability have generally focused on neighborhood- or precinct-level disparities, there are plausible crime benchmarks available. Researchers can study whether police appear to be treating minority

---

[166] This is why the hit-rate model has to be so complicated. One might have imagined a simpler statistical-discrimination story: police are rational Bayesians who interpret signals of "suspiciousness" in light of race-specific base rates and apply a lower suspiciousness bar for stopping one race than another. But whether the police have lowered the bar rationally—such that the *marginal* cases for each rate have the same hit rates—cannot be tested empirically by comparing *average* hit rates, which are calculated based on the whole group. Unfortunately, the data don't tell us which cases were on the margin. Knowles et al.'s model solves this problem by assuming hit rates equalize across all the other traits or behaviors that police observe—so at equilibrium, these traits tell police nothing, and all individuals are equally suspicious; all are marginal. *See* Knowles et al., *supra* note 156, at 209-12.

[167] It is often said, even by critics, that the model's main advantage is that it avoids omitted variable bias. *See, e.g.*, Childers, *supra* note 157, at 1033-35; Engel, *supra* note 40, at 16; Ayres, *supra*, at S79. This is a strange claim, because the model cannot empirically distinguish between racial discrimination and discrimination based on unobserved race-correlated traits; it assumes away the latter.

[168] *See* Coviello & Persico, *supra* note 106, at Appendix 11. This is a consequence of the equilibria assumed across all other traits; *see supra* note 166. When NYPD argues *both* that stop-and-frisk hit rates are equal across races, Floyd, 959 F.Supp.2d at 450-53, *and* that crime rates are not, *id.* at 584, it is contradicting itself. If both facts are true, the hit-rate model's assumptions are not. *See* Persico, *supra* note 104, at 1473 ("At equilibrium it cannot be that one group has a lower fraction of criminals.").

[169] *See* Banks, *supra* note 103, at 583; Harcourt, *supra*, at 1307-08. NYPD's stop-and-frisk program may be an exception; *see supra* note 87.

[170] Some suggest that the model's estimate of "taste-based" discrimination is a useful *lower bound* for unconstitutional discrimination. *E.g.*, Childers, *supra* note 157, at 1028. But the model's problematic assumptions and the omitted variable bias problem threaten the "lower bound" interpretation as well.

neighborhoods differently, and control for reported neighborhood crime and other neighborhood characteristics. Such studies cannot assess the effect of *individual* race on stops. But they represent the best current strategy for assessing whether neighborhoods are treated differently based on race.

A good example is the analysis of precinct-level disparities presented in the report of the plaintiffs' expert, Jeffrey Fagan, in the *Floyd* stop-and-frisk litigation.[171] Fagan's analysis regressed precinct stop rates on various racial groups' population share; controls included crime complaints the previous quarter, neighborhood socioeconomic and other demographic characteristics, and size of the precinct's police force.[172] The report found that black and Hispanic population share strongly predicted stop rates,[173] and the district court agreed.[174]

Crime rate controls in regressions could, in principle, could raise concerns like those explored in Part II. A neighborhood's stop rate incorporates stops of the guilty and the innocent, so we shouldn't expect it to be proportional to the neighborhood crime rate if *individuals* are stopped at equal rates across neighborhoods conditional on criminal conduct.[175] But our focus in this Part is whether the police are racially discriminating, not whether individuals with like conduct are being treated alike, so it makes sense to control for other factors that the police are taking into account. And when a department is making inter-neighborhood police allocation decisions, it very likely does consider neighborhood crime rates.[176]

Neighborhood regressions are affected by two other by-now-familiar problems: limited crime data and omitted variable bias. So researchers need to consider their crime benchmarks carefully, and should not claim to have "proven" discrimination or its absence decisively (although, again, results need not have a definitive causal interpretation to be useful).

3. Individual-Level Studies of Post-Stop Decisions

Some studies focus on disparities in searches, arrests, or other sanctions among stopped persons, controlling for other defendant and neighborhood characteristics and sometimes for the stop justification recorded by the officer.[177] Beyond the

---

[171] *Floyd v. City of New York,* 08 Civ. 1034, Report of Jeffrey Fagan 30-34.

[172] *Id.* The force size control means that the model does *not* test discrimination in allocation of police among precincts. An alternative model omitted this control, and estimated larger race gaps. *Id.* at 36.

[173] *Id.* at 32-34. The magnitudes suggest that moving from an entirely white neighborhood to an entirely black or Hispanic neighborhood would, other factors equal, nearly double the stop rate.

The plaintiffs' additional multilevel models assessed racial disparities within precincts as well. *Id.* at 40-45. These models do not have individual-level controls for criminal conduct or other possible individual-level confounders, which is a weakness from a causal inference perspective. On the other hand, if NYPD really was stopping people essentially at random, crime controls may have been unnecessary, for reasons discussed *supra* note 87.

[174] *Floyd*, 959 F.Supp.2d at 560.

[175] The math is not identical, because here the regressions include more variables and vary in functional form; many don't model the relationship between variables as a likelihood ratio.

[176] One cannot similarly defend *race-specific* neighborhood crime controls, which the NYPD argued for in certain analyses in *Floyd*. 959 F.Supp.2d at 584. Reliance on overall neighborhood crime rates is unconstitutional; reliance on race-specific group generalizations is not.

[177] E.g., Pickerill et al., *supra* note 40, at 9-19; Ridgeway & MacDonald, *supra* note 6, at 196; Greg

general omitted variable concern, the major limitations parallel those of studies of the post-arrest process, discussed in Part I.A: sample selection and post-treatment-control problems stemming from possible racial bias in the stop decision. When studying possible post-stop discrimination, one can't just assume that the stop decision by the same officer was unbiased. This is the mirror image of one of the problems discussed above with hit-rate studies, which instead assume that post-stop outcomes are objective measures of whether the stop was valid.

For example, consider Smith and Petrocelli's study of Richmond traffic stops. After controlling for officer and defendant traits and the stop location's crime rate, they found that among those stopped, minority drivers were substantially *less* likely to be ticketed or arrested.[178] What does that finding mean? The authors acknowledge the ambiguity, offering two possible interpretations: that the police avoided sanctioning minorities because they knew they were being studied, or that they sanctioned them less often because more of them had been unjustifiably stopped (an interpretation that shares the intuition of the hit-rate studies). Notice that these interpretations support opposite conclusions about the direction of discrimination. One solution to challenges like this is to combine analyses of post-stop outcomes with analyses of stop disparities, allowing estimates of unexplained post-stop disparities to be corrected for sample selection due to disparate stops.[179] But this is only viable when there is a sound method available for analyzing stops.

The literature on post-stop disparities also illustrates the importance and difficulty of choosing the right control variables. Again, we see more variations on a by-now familiar theme: researchers often estimate biases in police decisions while assuming that other police decisions are unbiased. For example, analyses of search rates often control for whether the individual is arrested or otherwise sanctioned.[180] The apparent rationale is that arrests and sanctions reflect conduct differences, and/or that searches incident to arrest are not discretionary. But arrest and sanction decisions are themselves discretionary (and thus potentially discriminatory).[181] Moreover, some arrests result from searches, not vice-versa, and some arrests may be motivated by the desire to carry out a search incident to arrest. When studying search decisions, it's inappropriate to control for something that's itself an outcome of the search decision; doing so likely biases disparity estimates downward.

More difficult dilemmas are posed by efforts to control for behavior as recorded

---

Ridgeway, *Assessing the Effect of Race Bias in Post-Traffic Stop Outcomes Using Propensity Scores*, 22 J. QUANT. CRIM. 1 (2006).

[178] Smith & Petrocelli, *supra* note 4, at 19-20.

[179] For example, two studies have compared traffic stop rates to violation rates measured via physical observation, then analyzed post-stop outcomes. *See* Engel et al., *supra* note 60, at 157 (finding that black and Hispanic drivers faced triple whites' search probability); Alpert et al., *supra* note 61, at 47 (finding no search-probability difference). Neither study's post-stop analysis corrected for sample selection, though in Alpert et al.'s study, this made sense because no stop disparity was found. Both studies' analyses of stop disparities are subject to the interpretive concerns raised in Part II.

[180] *See e.g.*, Alpert et al., *supra* note 61, at 47 (arrest); Pickerill et al., *supra* note 40, at 9-19 (citations).

[181] Likewise, it is misleading to refer to searches incident to arrest as "low-discretion" searches, *e.g.*, Pickerill et al., *supra* note 40, at 15, given that the arrest decision itself is highly discretionary.

by officers. For example, a Cleveland study found that officers more often described black drivers as noncompliant or disrespectful; arrest disparities disappeared after controlling for these descriptions.[182] The RAND study of NYPD's frisk, search, and sanction rates similarly controlled for "evasiveness, … appearing to be casing, acting as a lookout, wearing clothes consistent with those commonly used in crime, making furtive movements, acting in a manner consistent with a drug transaction or a violent crime, or having a suspicious bulge."[183] But officers' descriptions of these traits could be affected by race, or by the search or sanction decision itself (they could be post hoc justifications). These control variables could thus filter out part of what researchers are trying to measure. Other studies exclude such subjective factors from their models. I believe this is the better choice, but it does risk omitted variable bias if the descriptions reflect real differences.

As discussed above, a few self-report surveys (generally focused on youth) have included both criminal conduct and arrest questions, and these also ask a variety of other questions that relate to some potential confounding variables. It is possible to use these surveys to try to disentangle race's effects from those of other variables predicting arrest.[184] These studies are not affected by sample selection concerns stemming from the stop decision, because their samples are not confined to those stopped. However, they are still subject to concerns about omitted variables and about use of inappropriate controls. If a study uses only self-report information, it may omit factors relevant to police decisions but not known to the respondent (or asked about by the survey). Self-report data can sometimes be linked to official outcome data, but this does not really help with the omitted variable problem, because this official data won't cover people the police did not file reports on.

There are no perfect choices. In the best-case scenario, researchers will be able to make assumptions about selection bias and choice of control variables that have theoretical and/or empirical support and will investigate whether their estimates are robust to differing choices on difficult model specification questions. Careful observational studies of post-stop outcomes are potentially informative, but researchers must remember their limits.

4. Exploiting Variations in Enforcers' Information About Race

Some studies take advantage of variations in the information about race that is available to law enforcement. A couple have compared officers' traffic enforcement decisions to truly race-blind decisions: traffic-camera citations[185] and citations issued via aerial surveillance.[186] These are very informative designs, analogous to strong studies on other discrimination questions—for example, research demonstrating a

---

[182] Robin S. Engel et al., *Citizens' Demeanor, Race, and Traffic Stops*, in RACE, ETHNICITY, AND POLICING, *supra* note 6, at 297-99.

[183] Ridgeway, *supra* note 177, at 34-35.

[184] *See, e.g.*, Kirk, *supra* note 97.

[185] Montgomery County Department of Police, "Traffic Stop Data Analysis: Third Report," 2002.

[186] E.H. McConnell & A.R. Scheidegger, *Race and Speeding Citations: Comparing Speeding Citations Issued by Air Traffic Officers With Those Issued by Ground Traffic Officers*, Paper Presented at Annual Meeting of Academy of Criminal Justice Sciences, 2001.

spike in hiring of women when orchestras adopted blind auditions.[187] A limitation is that race information is not the only difference between the decision processes: automated and aerial enforcement target only particular unlawful behaviors, whereas human officers can observe various potential violations.[188] This problem parallels concerns about traffic benchmark studies that focus on a single speed cutoff.

Several traffic-stop studies have exploited variation in race information in circumstances that are arguably more similar: day and night. The studies compare stops at the same clock time but on either side of Daylight Savings Time transitions, such that they fall either just before or after nightfall. The intuition is that if higher black stop rates are driven by racial discrimination, the disparity should be reduced at night, when drivers' race is harder to see. Studies in Portland and Cincinnati found no reduction in disparity at night, concluding that disparities were not caused by racial discrimination.[189] Studies in Minneapolis and Syracuse reached the opposite conclusion; the Syracuse study, unlike the others, accounted for variations in artificial light.[190] These studies are very clever. But one interpretive problem is that darkness (not just clock time) might affect driver or police behavior through other channels, which the method cannot disentangle from the effect of reduced race information.[191]

Still, the general strategy of exploiting variations in race information is promising. However, its potential is limited to narrow contexts: those in which enforcement decisions can be made without close-range observation of suspects.

5. Lab Experiments on Implicit Biases

Aside from these observational approaches, many lab experiments demonstrate the prevalence of "implicit racial bias," including the association of blackness with criminality.[192] For example, Eberhardt et al. showed that police subjects who were primed subconsciously with crime-related images then paid disproportionate attention to black faces.[193] Crime-primed officers (bot not non-crime-primed officers) also strongly tended to pick the wrong black face out of a lineup—a more

---

[187] Claudia Goldin & Cecilia Rouse, *Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians*, 90 AM. ECON. REV. 715, 737-38 (2000).

[188] *See* Ridgeway & MacDonald, *supra* note 6, at 183 (citing these studies and raising this concern).

[189] Jeffrey Grogger & Greg Ridgeway, *Testing for Racial Profiling in Traffic Stops from Behind a Veil of Darkness*, 101 J. AM. STAT. ASSOC. 878 (2006); Terry Schell et al., Police-Community Relations in Cincinnati: Year Three Evalution Report, Rand Corp. Technical Report 535 (2007).

[190] Joseph A. Ritter & David Bael, *Detecting Racial Profiling in Minneapolis Traffic Stops*, CURA Reporter Spring/Summer 2009, at 11-17; William C. Horrace & Shawn M. Rohlin*, How Dark Is Dark? Bright Lights, Big City, Racial Profiling* (2014) (unpublished), http://www.colgate.edu/docs/default-source/d_academics_departments-and-programs_economics_colgate-hamilton-seminar-series/horracerohlindarkness-1-14-14.pdf?sfvrsn=0

[191] Darkness certainly affects driving behavior, and could also affect police tactics, or police perceptions of black criminality. In general, fear of crime is dramatically higher at night. *E.g.*, Kathleen A. Fox et al., *Gender, Crime Victimization, and Fear of Crime*, 22 SECURITY JOURNAL 24 (2009).

[192] *See, e.g.*, B. Keith Payne et al., *Weapon Bias*. 15 CURRENT DIR. IN PSYCH. SCIENCE 287 (2006) (reviewing literature); Quillian, *supra* note 143, at 314-20 (same); Kirwan Institute, Implicit Bias (2014), http://kirwaninstitute.osu.edu/wp-content/uploads/2014/03/2014-implicit-bias.pdf (same).

[193] Jennifer L. Eberhardt et al., *Seeing Black: Race, Crime, and Visual Processing*, 87 J PERSONALITY & SOCIAL PSYCH. 876, 885-88 (2004).

racially "stereotypical" black face.[194] A subset of this literature tests "shooter bias," using computer simulations; subjects are asked to "shoot" armed characters but not unarmed ones. These tests have found that players pick the right response faster if the image matches stereotypes (armed black or unarmed white characters).[195]

These studies are randomized experiments—the "gold standard" for causal inference. Many are quite small. But outside the lab, Internet-administered implicit bias tests have been taken by millions of people. Some of these test the association between blackness and weapons, which is prevalent: one analysis found that 72% of respondents showed this association, and only 9% showed the reverse.[196] Internet administration means test-taking conditions and samples are not controlled and respondents are not blind to the study's purpose. But people who choose to test themselves might actually be *less* biased than average, and if anything, most respondents are presumably trying to achieve an "unbiased" score. Moreover, tests that use subconscious primes and test quick reactions aren't easy to "game."

This research strongly indicates that implicit racial bias is fairly prevalent, including among police, but certainly not limited to them. Police and civilian subjects score similarly, and on some tasks they make fewer mistakes overall.[197] As Tonry puts it, given the bias found among "every imaginable group in the population, it would be remarkable if criminal justice practitioners were not affected."[198] Surveys have also shown widespread tendencies to *explicitly* associate blackness with criminality,[199] as well as overt endorsement of racial discrimination in other areas among a shrinking but still nontrivial subset of white respondents.[200]

The great unknown is how these phenomena translate into real-world decision-making by police.[201] While researchers have not yet linked implicit bias scores to real-world policing outcomes, such studies may be on the horizon. There are limitations to this approach, despite its promise: while the tests themselves are controlled experiments, using their results to explain real-world outcomes involves the usual

---

194 *Id.* at 887-88; *see also* Heather M Kleider et al., *Looking Like a Criminal*, 40 MEMORY & COGNITION 1200, 1200 (2012) (reaching similar findings with student subjects).

195 Joshua Correll et al., *The Police Officer's Dilemma: A Decade of Research on Racial Bias in the Decision to Shoot*, 8 SOC. & PERS. PSYCH. COMPASS 201, 206-07 (2014); Anthony G. Greenwald et al. *Targets of Discrimination: Effect of Race on Responses to Weapons Holders*, 39 J. EXPER. SOC. PSYCH. 399, 401-03 (2001).

196 Brian A. Nosek et al., *Pervasiveness and Correlates of Implicit Attitudes and Stereotypes*, 2007 EUR. REV. SOC. PSYCH. 1, 20.

197 Joshua Correll et al., *Across the Thin Blue Line: Police Officers and Racial Bias in the Decision to Shoot*, 92 J. PERSONALITY & SOCIAL PSYCH. 1006; *see generally* Correll et al., *supra* note 195 (reviewing literature).

198 Michael Tonry, *The Social, Psychological, and Political Causes of Racial Disparities in the American Criminal Justice System*, 39 CRIME & JUSTICE 273, 287 (2010).

199 James D. Unnever, *Race, Crime, and Public Opinion*, in OXFORD HANDBOOK OF ETHNICITY, CRIME, AND IMMIGRATION 70, 71 (Bucerius & Tonry eds.) (2014).

200 *See, e.g.*, Frank Newport, *In U.S., 87% Approve of Black-White Marriage, vs. 4% in 1958*, http://www.gallup.com/poll/163697/approve-marriage-blacks-whites.aspx (reporting 2013 poll showing that only 84% of white Americans approve of interracial marriage).

201 *E.g.*, BLANK ET AL., *supra* note 84, at 72 ("L]aboratory effects…can rarely tell us the extent to which naturally observed disparities are the result of discrimination.").

causal-inference challenges of observational research.[202] Moreover, the scores' explanatory value may understate race's total effects on officer decision-making because they test only specific subconscious mechanisms. Still, this research is a promising new line of inquiry into one plausible mechanism for disparities.

E. Testing Racial Profiling: The Promise of Auditing

Despite decades of effort, our current methods of evaluating police racial discrimination leave much unknown. I propose a new method to supplement the existing toolkit. "Auditing" refers to field studies that compare the treatment of paired "testers" who are similar but for a characteristic of interest. Such methods are used often in discrimination research and civil rights law enforcement in areas such as employment, housing, and lending. I propose to use testers (probably undercover police) to interact with police or to stage behavior that could attract their attention. Although it raises potential ethical, safety, legal, and political concerns, which I address here, this approach has substantial promise, capturing most of the advantages of lab experiments while directly testing real-world behavior.

1. Auditing in Research and Civil Rights Enforcement

A good example of the auditing approach is Ayres and Siegelman's study of race and sex discrimination by auto dealers.[203] The authors matched white male testers with black male, black female, and white female counterparts based on age, education, and assessed attractiveness; the testers all wore similar clothing and drove similar cars to the dealerships, where they negotiated prices on cars; black testers got substantially worse offers.[204] Other studies have used similar methods to study housing and employment markets,[205] plus various other phenomena—for example, a recent study found that drivers are less willing to yield to black jaywalkers.[206]

Instead of live testers, some studies manipulate only fictitious written information, such as employment applications,[207] student emails to professors,[208] and

---

[202] For example, an officer's experiences could influence both her IBT scores and her stop practices.

[203] Ian Ayres and Peter Siegelman, *Race and Gender Discrimination in Bargaining for a New Car*, 85 AM. ECON. REV. 304 (1995).

[204] *Id.* at 306, 319. The evidence of gender discrimination was less clear.

[205] *E.g.*, J. Yinger, *Measuring Discrimination with Fair Housing Audits*, 76 AM. ECON. REV. 881 (1986); *see* BLANK ET AL., *supra* note 84, at 106-07 (reviewing housing research); P.A. Riach & J. Rich, *Field Experiments of Discrimination in the Market Place*, 112 ECON. J. F480, F510-F513 (2002) same); Devah Pager, *The Use of Field Experiments for Studies of Employment Discrimination*, 609 ANNALS 104, 114 (reviewing employment research); Devah Pager, *The Mark of a Criminal Record*, 108 AM. J. SOC. 937 (2003) (studying effects of criminal records and race on employment).

[206] Tara Goddard et al., *Racial Bias in Driver Yielding Behavior at Crosswalks*, Portland State University, Working Paper, 2, http://ppms.otrec.us/media/project_files/TRF_Crosswalkpaper_Final.pdf; *see* BLANK ET AL., *supra* note 84, at 104-08 (reviewing auditing literature); Riach & Rich, *supra* note 205; Pager, *supra* note 205, at 113 tbl. 1 (same).

[207] *E.g.*, Marianne Bertrand & Sendhil Mullainathan, *Are Emily and Greg More Employable than Lakisha and Jamal?*, 94 AM. ECON. REV. 991 (2004); Pager, *supra* note 205, at 942-43 (reviewing studies).

[208] Katherine L. Milkman et al., *What Happens Before? A Field Experiment Exploring How Pay and Representation Differentially Shape Bias on the Pathway into Organizations* 24-26 (July 12, 2014) (unpublished manuscript), http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2063742.

writing samples.[209] Such designs allow true experimental manipulation of race and gender, which in-person tester studies don't quite achieve: one can randomize cases between testers, but one can't make the same tester white in one case and black in another. Instead, in-person auditing depends on careful matching and training to minimize within-pair variation.

No similar studies address U.S. law enforcement. In 1994, one criminal defendant introduced evidence from testers that he had hired to assess whether race affected Border Patrol stops. But the experiment was tiny and unscientific; the unpersuaded court observed that many relevant conditions had not been held constant.[210] A Mexico City study used testers who committed illegal left turns to test perceived-socioeconomic-status effects on police demands for bribes.[211] Another study focused not on police, but on private party suspicions of crime, testing store clerks' reactions to white and black shoppers.[212] An ABC News mini-experiment likewise tested private observers: actors cut the lock off a bicycle, and passerby reactions to the black actor were much more hostile.[213]

The use of testers is also a well-established civil rights enforcement strategy. In the 1950s, testers brought suits challenging public transit discrimination, and the Supreme Court upheld their standing.[214] Testers have played a prominent role in housing discrimination enforcement; the federal government has funded large tester studies and backed tester lawsuits brought by local fair housing associations.[215] Testers have also brought challenges to lending discrimination.[216] The Equal Employment Opportunity Commission has endorsed use of testers to challenge hiring discrimination,[217] though few cases have been brought.[218]

2. Auditing the Police: Key Research Design Considerations

Is auditing the police realistic? There are some good reasons that this hasn't been done before,[219] but I believe these can be addressed with careful research design.

---

[209] Arin N. Reeves, *Written in Black & White: Exploring Confirmation Bias in Racialized Perceptions of Writing Skills*, NEXTIONS 4-6 (Apr. 4, 2014).

[210] United States v. Beasley, 36 F.3d 1106, 1994 WL 504182, *4 (10th Cir. 1994) (unpublished).

[211] Brian J. Fried et al., *Corruption and Inequality at the Crossroad*, 45 LATIN AM. RES. REV. 76 (2010).

[212] George E. Schreer et. al, *"Shopping While Black": Examining Racial Discrimination in a Retail Setting*, 39 J. APPLIED SOC. PSYCHOL. 1432 (2009).

[213] ABC, *What Would You Do? (Bike Thief),* http://www.youtube.com/watch?v=S0kV_b3IK9M.

[214] *Evers v. Dwyer*, 358 U.S. 202 (1958).

[215] *See* Havens Realty Corp. v. Coleman, 455 U.S. 363 (1982) (upholding tester standing); Michael Selmi, *Public vs. Private Enforcement of Civil Rights*, 45 U.C.L.A. L. REV. 1401, 1426 (1998); M.A. Turner et al., *Discrimination in Metropolitan Housing Markets: National Results from Phase I HDS 2000* (2002).

[216] Steve Tomkowiak, *Using Testing Evidence in Mortgage Lending Discrimination Cases*, 41 URB. LAW. 319, 326-336 (2009).

[217] Equal Employment Opportunity Comm., Dec. No. N-915-062, Policy Guidance on the Use of EEO Testers, Nov. 20, 1990; Julie Lee & Caitlin Liu, *Measuring Discrimination in the Workplace*, 6 U. CHI. L. SCH. ROUNDTABLE 195, 213 (1999).

[218] Marc Bendick, Jr. & Ana P. Nunes, *Developing the Research Basis for Controlling Bias in Hiring*, 68 J. SOC. ISSUES 238, 256 (2012).

[219] Indeed, aside from the *Beasley* defendant's effort, *see supra* note 210, it has hardly been suggested. One scholarly piece and one news article each give the idea a sentence or two. Pamela S. Karlan, *Race,*

Here, I address several objectives that researchers must balance: safety, legality, importance, methodological rigor, statistical power, and cost concerns.

Safety. A paramount concern is minimizing risk to testers, police, and third parties. The research designs I propose below involve no serious law-breaking, nor do they suggest a violent situation. They are not designed to test arrest probability, but to potentially elicit relatively minimal police interactions. Testers must be trained to be absolutely cooperative. The safest approach would involve law enforcement participation: voluntary or court-ordered police department self-monitoring or outside civil-rights agency investigations. Ideally, testers could be undercover agents—people who regularly carry out far riskier work than this—and police backup could be ready to intervene if any safety threat arises.

The designs proposed below also pose minimal risk to the officers being studied. With just one or two interactions with each officer, they would be used to diagnose broad patterns, not to identify individual "bad apples." They also involve very minimal officer time, minimizing distraction from ordinary public-safety duties.

Legality. The criminal law constrains staging of actual crimes, lying to the police, and recording of interactions.[220] This is another advantage of governmental involvement. Undercover police routinely participate in otherwise-criminal activity and enjoy effective immunity from prosecution.[221] Private testers can't be asked to commit serious crimes, but might choose to risk minor violations, as did researchers in several studies mentioned above: Lamberth's Turnpike study, the jaywalking study, and the Mexico City bribery study. Most of the designs proposed below involve no lawbreaking or lying, just potentially suspicious activity.

Importance. Studies should focus on contexts in which there is reason to suspect discrimination (for example, large raw disparities, or citizen complaints) and in which discrimination would have meaningful consequences. But such contexts need not involve serious crimes. Misdemeanor enforcement can result in detention and substantial collateral consequences, can be highly stressful, may be a pretext to look for more serious criminality,[222] and may be a method of expanding the surveillance "net," exposing arrestees to more police interactions in the future.[223]

Methodological rigor. The most obvious requirement for effective auditing is that the deception work. The interaction should thus be quite ordinary, brief, and

---

Rights, and Remedies in Criminal Adjudication, 96 MICH. L. REV. 2001, 2008 (1998); Emily Badger, *Why It's So Hard to Study Racial Profiling By Police*, WASH. POST, April 30, 2014. Reviews of methods for studying racial profiling omit it; for example, Blank et al. don't mention auditing in their chapter on police, *supra* note 84, at 186-204, even though they endorse it for other contexts like housing, *id.* at 103-117.

[220] In most states, anyone may record their own interactions without permission, though some states require two-party consent. Reporters Committee for the Freedom of the Press, Reporter's Recording Guide 2-3, http://www.rcfp.org/rcfp/orders/docs/RECORDING.pdf (2010). There may also be a constitutional right to record police, *e.g.*, Glik v. Cunniffe, 655 F.3d 78, 82-84 (1st Cir. 2011), though some courts require *open* recording, Crawford v. Geiger, 996 F.Supp.2d 603, 614 (2014).

[221] Elizabeth E. Joh, *Breaking the Law to Enforce It: Undercover Police Participation in Crime*, 62 STAN. L. REV. 155, 157, 165-69 (2009).

[222] *See supra* note 121 (discussing *Whren*). Arrestees may be searched without warrants.

[223] Issa Kohler-Hausmann, *Managerial Justice and Mass Misdemeanors*, 66 STAN. L. REV. 611, 632-33 (2014).

forgettable. Observations should be distributed across different police beats and shifts and across time, so that individual officers are unlikely to notice patterns.

The primary threat to causal inference from auditing studies is tester heterogeneity,[224] so testers should be matched carefully.[225] Even so, subtle differences may remain, but training combined with simple, easy-to-replicate "scripts" can make these less likely to affect outcomes. Analyses could focus on outcomes, like whether any interaction occurs, that are unaffected by subtle differences in conversational styles. Optimally, the testers should be blind to the study's purpose (for example, they could be told they are testing enforcement without mentioning the racial dimension),[226] though this might be hard to pull off. But testers' activities could be recorded and later coded by persons who are blind to the purpose.

One possible interpretive challenge is discerning whether racial differences in police actions might result from disparities in citizens' calls to the police, rather than police discrimination. With police department cooperation, this mechanism could be teased out, because the police could collect information on citizen calls.

Statistical power and cost. The sample size must provide sufficient statistical power to produce reasonably precise estimates.[227] Ideally, this means at least hundreds of observations[228]--a plausible number (large cities have thousands of officers), provided the tests are spread across beats and shifts.[229] Many published auditing studies have much smaller samples, allowing them only to detect large effects, and even then, imprecisely.[230] Although larger samples produce greater power, they cost more, and may increase the risk of police noticing patterns. This is another reason to use designs that involve low-intensity, brief, forgettable interactions—they can be repeated more often at reasonable cost. However, the interactions do need to be designed to elicit police responses reasonably often; to heighten the chance of such response, testers should be positioned near the known location of officers.

3. Possible Research Designs

Here, I list a few examples of research designs, leaving the details to be tailored to the city and police force.

---

[224] *See, e.g.*, James J. Heckman, *Detecting Discrimination*, 12 J. ECON. PERSP. 101, 108-09 (1998).

[225] *See* Pager, *supra* note 205, at 111-12, 123-24. But researchers should avoid too-perfect matches on traits that themselves signify race (*e.g.*, hair). *See* Riach & Rich, *supra* note 205, at F483-F484.

[226] *See* Ayres & Siegelman, *supra* note 203 (using blind testers); Lee & Liu, *supra* note 217, at 224.

[227] Power analyses are typically framed in terms of hypothesis-testing, wherein power is the probability of obtaining a statistically significant result if the "true" effect is of a certain size. Power depends on sample size, the size of effect one seeks to detect, the statistical significance threshold, and (for binary outcomes) the baseline frequency of the outcomes.

[228] Sample-size calculators are widely available; they require assumptions about effect size. For example, if one seeks 80% power with a 95% confidence level, assuming the true probabilities for the two groups are 30% and 40% respectively, a common power formula requires a total sample size of 708. *See* "Power (Sample Size) Calculators," https://www.sealedenvelope.com/power/binary-superiority/. If the probabilities were 30% and 50%, the sample size required would be smaller (182).

[229] For example, the Chicago police department has 279 distinct beats, each patrolled by eight or nine officers. Chicago Police Department, "Beat Officers," https://portal.chicagopolice.org/portal/page/portal/ClearPath/Get%20Involved/How%20CAPS%20works/Beat%20Officers.

[230] *E.g.*, Fried et al., *supra* note 211 (43 tests); Schreer et al, *supra* note 212, at 1438 (31 tests, 6 stores).

Open Container/Minor in Possession. Testers could walk past beat officers carrying a container of liquid, such as a soda bottle that resembles a beer bottle, testing whether they're asked what's in it. If the containers do *not* actually contain alcohol, suspicion could be immediately dispelled.

Loitering. Testers, in same-race pairs, could hang out in public, testing whether police approach. To increase rates of police interactions, testers could engage in further "nuisance" activity, like playing music or smoking, or wear bulky clothing.

Casing. Testers could wait outside jewelry or other stores, looking in—behavior that could be construed either as "window-shopping" or "casing."

Bike or car theft. Testers could break a bike lock or break into a car using a coat hanger—like the ABC News video described above, but larger-scale. In the car example, testers could carry the registration so as to dispel suspicion quickly. A challenge will be objectively differentiating hostile interactions from offers to help.

Traffic violations. Testers could break traffic laws and see if they get stopped (and searched). While safety would be a concern, some traffic violations could pose little or no danger—for example, expired or missing license plates.

Checkpoints. Checkpoints are promising settings for auditing: some law enforcement contact is guaranteed, the location is fixed, the setting is highly monitored and low-risk, and the testers' activity (just passing through) would be unremarkable. Outcomes could include time elapsed and diversion for extra searches. Agency cooperation, while not essential, would help; it would allow access to the information agents obtain when they run individuals' identification.

Manipulation of Victim Reports, Police Files, and Training Exercises Other strategies could avoid in-person police encounters. "Victims" (perhaps themselves of varied race) could call in crime reports with varied suspect race, to test differences in dispatchers' response (assuming a mechanism is in place to quickly cancel the investigation). Race could be manipulated in training exercises involving assessment of case files or descriptions. Manipulation of police files could also be used to test prosecutors' charging or intake decisions.

Responding to Citizen Complaints. Officers that staff citizen outreach or internal affairs departments could be tested to see if they respond differently to complaints about officers depending on the complainant's race.[231] The test should focus on initial intake, with a mechanism for stopping the ensuing investigations.

4. Advantages and Limitations

In real life, race mediates the lives people lead, but auditing measures disparate treatment of individuals who are doing the same thing in the same places. This is both a strength and a limitation. On the one hand, it enables sound causal inferences: if we eliminate differences other than race, we can more confidently attribute disparate outcomes to racial discrimination. Auditing designs would be much better

[231] *See, e.g.*, New York Comm'n to Combat Police Corruption, *Follow-Up Review of the Internal Affairs Bureau Command Center* 1-5, 17 (1999) (describing center that takes 20,000 complaints per year). *See also* Douglas S. Massey & Garvey Lundy, *Use of Black English and Racial Discrimination in Urban Housing Markets*, 36 URB. AFF. REV. 452, 456-59 (2001) (discussing phone-based auditing studies).

tailored to isolate the effects of racial discrimination than regression studies and other observational approaches. If testers are matched and trained well, it could approximate a true experiment, but in a real-life setting, not a lab.[232]

But auditing may miss dimensions of real-world racial discrimination. For example, if the police heavily target young men who dress a certain way, and virtually all such young men are black, perhaps clothing style is not a confounder that should be filtered out via the use of identically dressed testers, but rather a race proxy—a mechanism for racially disparate treatment. Similarly, most of the designs above would test disparities *within neighborhoods* (or at checkpoints), and would miss differences driven by neighborhood racial composition.

The auditing design could, however, be extended to test the effects of such race-correlated variables and their interaction with race—for example, by changing the same testers' clothing and/or sending them to different neighborhoods. An advantage over observational studies of inter-neighborhood disparities is that this approach could rule out inter-neighborhood differences in individuals' behavior, although it would not necessarily rule out all other neighborhood differences. Similarly, evidence that the police disfavor some characteristic like a clothing style would not definitively prove that they are using it as a race proxy.

Auditing would produce context-specific estimates, not an overall measure of racial discrimination in stops or arrests.[233] These estimates will be more informative if the test is similar to some class of activity that produces a reasonable share of the department's stops or arrests. Loitering and minor-in-possession are good examples.

5. Implementation

Given its longstanding role in civil rights enforcement, federal or state agencies' use of auditing to assess police disparities is plausible. Tester programs in other areas have sometimes been controversial,[234] and may well be in this context as well, but there are countervailing political pressures. In surveys, large majorities oppose racial profiling.[235] DOJ's Civil Rights Division has a strong interest in the issue and in police abuses generally,[236] and the issue has been an especially high overall DOJ priority in the wake of the Ferguson shooting.[237]

---

[232] *See* Quillian, *supra* note 143, at 303 ("[A]udit studies often are the best method for measuring…discrimination.").

[233] *Cf.* Heckman, *supra* note 224, at 102-11 (criticizing employment audit studies for not providing estimates of market discrimination).

[234] *See* Selmi, *supra* note 215, at 1427; Alex Young K. Oh, *Using Employment Testers to Detect Discrimination*, 7 GEO. J. LEGAL ETHICS 1473, 480 (1993) (citing employer fears of tester litigation).

[235] Emily Eakins, *Poll: 70% of Americans Oppose Racial Profiling by the Police*, Reason-Rupe Poll, Oct. 14, 2014, http://reason.com/poll/2014/10/14/poll-70-of-americans-oppose-racial-profi.

[236] *Addressing Police Misconduct Laws Enforced by the Department of Justice*, U.S. DEP'T OF JUSTICE, http://www.justice.gov/crt/about/spl/documents/polmis.php; U.S. DEP'T OF JUSTICE, CIVIL RIGHTS DIV., GUIDANCE REGARDING THE USE OF RACE BY FEDERAL LAW ENFORCEMENT AGENCIES (2003), http://www.justice.gov/crt/about/spl/documents/guidance_on_race.pdf.

[237] Statement of Attorney General Eric Holder, Latest Developments in Federal Civil Rights Investigation in Ferguson, MO (Aug. 14, 2014).

Outside-agency auditing would lose some of the advantages of police-department self-monitoring (for example, access to internal data), but it would otherwise retain the advantages of being able to use trained undercover officers and protect them from physical or legal harm. The outside-enforcement approach would face less risk of being compromised by leaks or internal resistance. It is the most plausible strategy when a police department is hostile to scrutiny. Auditing could also be required by court order or settlement in civil rights litigation. Analogously, a major benchmarking study was carried out by the New Jersey Attorney General's office pursuant to a settlement with DOJ,[238] and outside monitors have been appointed for numerous police departments under consent decrees.[239]

Voluntary self-auditing by police departments is promising, but is it realistic? After all, adverse findings could be embarrassing and invite litigation. Moreover, the studies could be resource-intensive and risk angering officers and unions or even the undercover testers themselves. Still, while many departments would doubtless reject the idea, the 18,000 law enforcement agencies in the U.S. are not monolithic, and there may well be some who embrace the idea. Typically, agency heads are political appointees, and there is no reason to assume that all cities' political leaders would be primarily interested in hiding racial discrimination, rather than eliminating it.

Hundreds of police departments have already invested considerable resources in collecting racial disparity data, and many have carried out ambitious studies.[240] Some police departments have "early warning" programs to identify individual problem officers.[241] Any of these programs risks litigation or officer backlash—indeed, programs that risk getting individual officers in trouble may raise a worse risk of backlash than auditing does.[242] These risks have not precluded their adoption.

There is substantial precedent for undercover police work to help departments self-diagnose problems. Some departments use a practice called "red teaming" to test police responses to security threats and emergency situations.[243] Undercover agents are also often employed in police corruption investigations.[244] Several police

---

[238] *See* Lange et al, *supra* note 205, at 196-97.

[239] *Floyd v. City of New York*, 08-CIV-1034, Statement of Interest of the United States, June 2, 2013 (advocating court appointment of monitor); Barbara Attard, *Oversight of Law Enforcement Is Beneficial and Needed—Both Inside and Out*, 30 PACE L. REV.1548, 1550 (2010).

[240] *See, e.g.,* Engel et al., *supra* note 60; Ctr. for Policing Equity, *What We've Done*, http://cpe.psych.ucla.edu/what-weve-done (describing CPE's work with police departments).

[241] Engel & Calnon, *supra* note 41, at 109; *see* Ridgeway, *supra*, at 21-30.

[242] Unions generally strongly oppose policies with potential adverse consequences for individual officers. Engel & Calnon, *supra* note 41, at 109; Kevin M. Keenan & Samuel Walker, *An Impediment to Police Officer Accountability?*, 14 B.U. PUB. INT. L.J. 185, 198-99 (2005).

[243] The term comes from military wargaming exercises. Michael K. Meehan, *Red Teaming for Law Enforcement*, POLICE CHIEF**;** *see* Fed. Bureau of Investigation, Subject Bibliography: Red Teaming, http://fbilibrary.fbiacademy.edu/bibliographies/redteaming.pdf (collecting sources); William H. Adcox, *The Red Team: An Innovative Quality Control Practice in Facility Security*, 74 POLICE CHIEF (2007) (describing "breach exercises" carried out by undercover teams at protected facilities).

[244] *E.g.*, Steve Rothlein, *Conducting Integrity Tests on Law Enforcment Officers*, Legal Liability and Risk Management Institute, http://www.llrmi.com/articles/legal_update/le_integrity_tests.shtml (2010); *see* Tim Prenzler & Carol Ronken, *Police Integrity Testing in Australia*, 1 CRIMINOLOGY & CRIM. J. 319,

departments (including New York and Los Angeles) regularly conduct "random integrity tests"—exposing officers to random stings.[245] Corruption is likely as embarrassing to police departments as racial discrimination is—yet these departments have carried out the corruption equivalent of auditing.

But even if departments can be persuaded to undertake auditing studies, can they be trusted not to undermine their accuracy? Internal affairs divisions and police leadership have often been sharply criticized for papering over police misconduct and corruption.[246] Under the right conditions, however, the prospects for effectiveness are reasonable. Self-studies will be more credible if undertaken together with outside watchdog organizations or academic researchers who have control over data collection and analysis[247]--provided those outside actors are truly independent.[248] Undercover agents, presumably borrowed from other departments, would have to be carefully chosen, because they would have to be trusted not to tip off other officers or to try to manipulate the study's findings.[249]

If police departments are reluctant to expose themselves to criticism and liability, or to anger their own officers, they could conduct internal auditing programs without publicizing results, or ask academic collaborators to publish anonymized results. To encourage self-studies, legislatures could consider enacting statutory evidentiary privileges. Congress has enacted just such "self-testing" privileges for mortgage lenders and creditors in the Fair Housing Act and the Equal Credit Opportunity Act.[250] These apply only if, upon discovering evidence of discrimination, the lender undertakes "appropriate corrective action."[251] If legislatures applied similar privileges to police self-testing, they would be modest extensions of the "self-criticism privileges" that law enforcement agencies already often invoke (which cover subjective analyses but not underlying facts).[252]

---

319 (2001) (describing undercover integrity testing in Australia as an "essential" anticorruption tool).

[245] Rothlein, *supra* note 244; Prenzler & Ronken, *supra* note 244, at 321-23; Sanja Kunjak Ivkovic, 93 J. CRIM. L. & CRIMINOLOGY 593, 617-19 (2003).

[246] *E.g.*, Ivkovic, *supra* note 245, at 596-97.

[247] This is the modus operandi of the Center for Policing Equity, which connects researchers with police departments. Center for Policing Equity, *What We Do*, http://cpe.psych.ucla.edu/. *See* Merrick Bobb, *Civilian Oversight of the Police in the United States*, 22 ST. LOUIS UNIV. L. REV. 151, 159-63 (2003) (describing some departments' voluntary use of accountability organizations, independent investigators, and civilian review boards to monitor use of force and corruption).

[248] Civilian oversight boards have often been criticized for being overly deferential to police. *E.g.*, Stephen Clarke, Note, *Arrested Oversight: A Comparative Study of How Civilian Oversight of the Police Should Function and How it Fails*, 43 COLUM. J.L. & SOC. PROBS. 1, 11-12 (2009). Academic researchers with external (non-police) funding may be better equipped to provide accountability, but it will be important to negotiate agreements preserving researchers' control over reporting of results.

[249] Riach & Rich, *supra* note 205, at F483, worry that "consciously or unconsciously, minority applicants may be motivated to prove the existence of discrimination." *See* Heckman, *supra* note 224, at 104. When police are investigating police, one might worry more about the opposite concern.

[250] *See* Tomkowiak, *supra* note 216, at 326-27.

[251] *Id.*; *see* ADI Consulting, *The Self-Testing Privilege for Fair Lending Compliance*, http://www.adiconsulting.com/Docs/2006%20FL%20Self-Testing%20Privilege.pdf (also describing some similar privileges developed by states).

[252] *See* Josh Jones, Note, *Behind the Shield? Law Enforcement Agencies and the Self-Critical Analysis Privilege*,

If government involvement proves impracticable, academic researchers might be able to carry out some of the designs on their own. Academic research is by Institutional Review Board oversight,[253] but IRBs generally focus on harms to subjects (here, police) and perhaps third parties. Here, essentially all the risk is on the research staff (the testers).[254] But while the IRB may not regulate such risks, ethical researchers should consider them. While well-informed research staff should be free to take on non-zero risks (as much research does), supervisors should aim to keep this risk minimal, especially if they are students, who may be reluctant to refuse.

Overall, while auditing designs could face serious practical and political hurdles, their use is plausible. They offer a potentially valuable new addition to the toolkit of researchers, civil rights agencies, and police departments.

CONCLUSION

There is no single correct empirical method for assessing racial disparities in policing, because there is no single correct normative conception of what kind of inequality we should care about. I have focused my analysis principally on how to get at two estimands that I consider especially important: first, racial disparity in police interactions conditional on criminal conduct, and second, the effects of police racial discrimination. The policing-crime comparisons that have dominated the literature give a skewed sense of the first, generally overcorrecting for crime differences. Meanwhile, several available methods for assessing police racial discrimination are useful, but are limited by various causal-inference and external-validity concerns; auditing, despite its challenges, is an appealing alternative.

Some will no doubt disagree with me as to what the most important empirical questions are. I hope that this discussion might persuade those with different views to articulate them clearly and to employ statistical analyses and numerical comparisons that are consistent with their normative premises. An empirical analysis that is well designed to answer one question may produce dramatically wrong conclusions if it is misinterpreted as an answer to another. The conceptual confusions that have pervaded public and scholarly debates about race and policing are not merely a matter of laypeople misunderstanding statistics or activists strategically misusing them. One can hardly expect more out of the public debate when much of the underlying empirical literature is itself riddled with the same problems. We can and should do better.

---

60 WASH. & LEE LAW REVIEW 1609, 1611-14 (2003). Federal privilege legislation could be grounded in Congress's Fourteenth Amendment enforcement powers, and could perhaps extend to state courts.

[253] This may be true even if researchers work with government, depending on their roles.

[254] Research guidelines also generally permit dispensing with informed consent if the research design requires it (as it does here) and the potential harm is minimal. *See* Pager, *supra* note 205, at 126.

APPENDIX

Tables

Table A1. Examples Assuming Racial Equality Conditional on Conduct

	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]
Guilty Stop Rate: S_G/P_G	50%	25%	50%	50%	50%	50%	50%	50%	50%
Innocent Stop Rate: S_I/P_I	25%	25%	5%	5%	5%	25%	25%	25%	25%
Black Guilt Rate: P_{GB}/P_B	40%	40%	40%	8%	8%	96%	40%	40%	20%
White Guilt Rate: P_{GW}/P_W	20%	20%	20%	4%	2%	92%	20%	20%	20%
Num. Black: P_B	1000	1000	1000	1000	1000	1000	1000	4000	1000
Num. White: P_W	1000	1000	1000	1000	1000	1000	4000	1000	1000
Stop Ratio/Crime Ratio $\frac{(S_B/S_W)}{(P_{GB}/P_{GW})}$	0.58	0.5	0.82	0.63	0.36	0.98	0.58	0.58	1
Stop Share/Crime Share $\frac{(S_B/S)}{(P_{GB}/P_G)}$	0.81	0.75	0.93	0.84	0.74	0.99	0.68	0.93	1
Hit Rate Ratio $\frac{(S_{GB}/S_B)}{(S_{GW}/S_W)}$	1.71	2	1.22	1.58	2.74	1.02	1.71	1.71	1

General notes:

a. All scenarios apply the same stop rates to people with the same criminal conduct, regardless of race. The first six lines show hypothetical assumptions; the bottom three lines show ratios calculated based on those assumptions. All assume higher black crime rates (except [9], which assumes no difference), paralleling the usual argument made in defense of policing disparities.

b. Although all the stop rates in these examples are higher than one would find in most real-world policing contexts, one could divide every stop rate by 10 (or by anything) and it would not affect any of the ratios.

c. The Hit Rate Ratio is discussed in Part II.D. It is 1 if stops are equally productive *(likely to succeed in catching a criminal) across racial groups. Note that here, policing is racially equitable conditional on conduct, but (except in Column 9, where there is no crime difference) hit rates are never aligned.*

Notes on each column:

[1] parallels the Table 1 example in the text.

[2] and [3] vary the degree to which the police accurately discern guilt (i.e., varies the gap between innocent and guilty stop rates). The Ratio/Ratio and Share/Share measures are less misleading when the police are more discerning, because there are fewer stops of the innocent.

[4] lowers guilt rates for both races. Although (as in [3]) the police are ten times as likely to stop the guilty, 61% of stops overall are of the innocent, and the Ratio/Ratio and Share/Share measures perform badly.

[5] further lowers the white guilt rate. With a larger crime-rate difference, the measures are even more misleading.

[6] is the same as [1] except guilt rates are very high. With few innocents, and thus few stops of the innocent, the Ratio/Ratio and Share/Share measures are close to the true ratio of stops among the guilty.

[7] and [8] are the same as [1] except that the relative population sizes vary. Only the Share/Share measure changes as a result. It is always closer to 1 than the Ratio/Ratio measure, and is more "diluted" in this sense the larger the high-crime group's population share is.

[9] is the same as [1] except it eliminates the crime-rate disparity. The problems with the Ratio/Ratio and Share/Share measures disappear; they emerge only when there are crime-rate differences.

Table A2. Examples Assuming Racial Disparity Conditional on Conduct

	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]
Black Guilty Stop %: S_{GB}/P_{GB}	50%	50%	40%	50%	40%	40%	40%	50%
White Guilty Stop %: S_{GW}/P_{GW}	25%	20%	30%	50%	70%	70%	70%	25%
Black Innocent Stop %: S_{IB}/P_{IB}	30%	20%	25%	50%	20%	20%	20%	30%
White Innocent Stop %: S_{IW}/P_{IW}	15%	15%	10%	10%	35%	35%	35%	15%
Black Guilt %: P_{GB}/P_B	40%	20%	20%	20%	20%	20%	20%	20%
White Guilt %: P_{GW}/P_W	20%	10%	10%	10%	10%	10%	10%	20%
Number Black: P_B	1000	1000	1000	1000	1000	1000	4000	1000
Number White: P_W	1000	1000	1000	1000	1000	4000	1000	1000
Stop Rate Ratio: Guilty (D_G) $(S_{GB}/ S_{GW})/ (P_{GB} /P_{GW})$	1.67	2.5	1.33	1	0.57	0.57	0.57	2
Stop Rate Ratio: Innocent (D_I) $(S_{IB}/ S_{IW})/ (P_{IB} /P_{IW})$	1.67	1.33	2.5	5	0.57	0.57	0.57	2
Stop Ratio/Crime Ratio $(S_B/ S_W)/ (P_{GB} /P_{GW})$	0.97	0.84	1.17	1.79	0.31	0.31	0.31	2
Stop Share/Crime Share $(S_B/ S)/ (P_{GB} /P_G)$	0.99	0.94	1.05	1.17	0.58	0.40	0.80	1.33
Average Stop Rate Ratio, Holding Guilt Constant at Black Mean (Weighted Mean of D_G and D_I)	1.67	1.63	2	2.78	0.57	0.57	0.57	2
Hit Rate Ratio $(S_{GB}/S_B)/(S_{GW}/S_W)$	1.72	2.98	1.14	0.56	1.83	1.83	1.83	1

General Notes
a. The first eight lines are assumptions; the last six are calculated ratios that represent different possible measures of disparity. In addition to the measures from Table 1A, this table includes D_I, D_G, and a weighted average of the two called the "Average Stop Rate Ratio Holding Guilt Constant at Black Mean." These measures were not included in Table 1A because there we were assuming no disparities in stop rates conditional on conduct (so all of them would have been 1).
b. The Average Stop Rate Ratio line represents one of several reasonable ways of averaging the stop rate ratio among the guilty and the stop rate ratio among the innocent to produce an overall average stop rate ratio conditional on criminal conduct. It reflects a reweighting exercise that compares an observed outcome to a counterfactual outcome. It asks: By what proportion does the number of black stops differ from the number that we would have seen if the white *conditional stop probabilities had been applied to the black population? This proportion can be expressed as:*

$$\text{Average Stop Rate Ratio} = \frac{S_B}{P_{IB} * {}^{S_{IW}}/_{P_{IW}} + P_{GB} * {}^{S_{GW}}/_{P_{GW}}}$$

This proportion represents the multiplicative effect on stops of applying the black conditional stop probabilities (instead of the counterfactual white ones) to the black *population—or, put another way, it is the average disparity holding guilt rates constant at the black average. In the notes on Proof 3, I discuss two other, similar ways one might reasonably estimate "average disparity" conditional on conduct, holding guilt rates constant either at the white average or at the overall population average. All three versions always produce average stop rate ratios that fall between D_I and D_G.*

Notes on each column
[1] parallels the Table 2 example in the text.

[2] and [3] likewise assume that black pedestrians are much more likely to be stopped conditional on criminal conduct, but here disparities are not uniform across criminal conduct conditions. The disparity is larger among the guilty in [2] and the innocent in [3]. The Ratio/Ratio *and Share/Share measures again mask these disparities, or (in [2]) reverse their apparent direction. One might expect disparities that are similar across conduct conditions (as in [1]) if police are paying more attention across the board to black pedestrians, or have stationed more officers in their neighborhoods. One might expect greater disparity among the guilty (as in [2]) if the police both are paying more attention to black pedestrians and are more accurate in discerning their guilt. One might expect greater disparity among the innocent (as in [1]) if the police lower the suspicious-behavior bar for stopping black pedestrians, affecting mostly the innocent.*

[4] again assumes racial disparity conditional on conduct, disfavoring blacks, but this time it is entirely driven by a large disparity in stops of the guilty. In scenarios with dissimilar disparities among the innocent and the guilty, it is possible for the Ratio/Ratio *and Share/Share measures to be higher than D_I but lower than D_G (or vice-versa), whereas in most of the examples they are lower than both. (The opposite scenario would also be possible.) However, the* Ratio/Ratio *measure systematically misleads in a particular direction in the sense that it always appears to be more favorable to the higher-crime group than the true* average *stop rate ratio conditional on criminal conduct.*

[5] through [7] show examples in which we assume the true disparity cuts the other direction: white stop rates are higher conditional on criminal conduct. The Ratio/Ratio *measure now substantially* exaggerates *the disparity, making it look like white pedestrians are only 31% as likely to be stopped as black pedestrians, conditional on criminal conduct (when the actual assumed ratio is 57%). The only difference between these three columns is the group population sizes, which affect the Share/Share measure only. When true disparities conditional on criminal conduct cut in* favor *of the higher-crime group, the Share/Share measure does not always mislead in the same direction, because of its "dilution" toward 1 (relative to the* Ratio/Ratio *measure), which is more pronounced when the group whose shares are being compared (here, blacks) is a larger share of the population.*

[8] shows an example that parallels [1], except with equal black and white crime rates. The Ratio/Ratio *measure is no longer misleading—it matches the actual average stop rate ratio. The Share/Share ratio is closer to 1, as in all examples.*

Proofs

Proof 1.

This proof shows that when there is racial equality in policing rates conditional on criminal conduct, but crime rates differ and not all stops are of the guilty, the higher-crime group always appears relatively "underpoliced" according to the Stop Ratio/Crime Ratio and Stop Share/Crime Share measures. The proof begins by showing that the Ratio/Ratio measure is less than 1 under these conditions (with the higher-crime group in the numerator), and proceeds to show that the Share/Share measure is also less than 1.

Definitions:

	Group 1	Group 2	Combined
Populations	$P_1 = P_{I1} + P_{G1}$	$P_2 = P_{I2} + P_{G2}$	$P = P_1 + P_2$
Stops	$S_1 = S_{I1} + S_{G1}$	$S_2 = S_{I2} + S_{G2}$	$S = S_1 + S_2$

*Subscripts I and G denote innocent and guilty subsets.

Assumptions:

(a)	Group 1 has a higher crime rate.	$\frac{P_{G2}}{P_2} < \frac{P_{G1}}{P_1}$
(b)	Equal stop rates of innocent.	$D_I = \frac{S_{I1}/P_{I1}}{S_{I2}/P_{I2}} = 1$
(c)	Equal stop rates of guilty.	$D_G = \frac{S_{G1}/P_{G1}}{S_{G2}/P_{G2}} = 1$
(d)	For both groups, P_I, P_G, S_I, and S_G are > 0.	

Objective 1. Prove: $\frac{S_1/S_2}{P_{G1}/P_{G2}} < 1$

(1) Apply Population definitions to (a). $\frac{P_{G2}}{P_{I2} + P_{G2}} < \frac{P_{G1}}{P_{I1} + P_{G1}}$

(2) Rearrange terms. $\frac{P_{I1} + P_{G1}}{P_{G1}} < \frac{P_{I2} + P_{G2}}{P_{G2}}$

(3) Simplify (subtract 1 from both sides). $\frac{P_{I1}}{P_{G1}} < \frac{P_{I2}}{P_{G2}}$

(4) Divide by equivalent terms. [From (b), we know $P_{I1}S_{I2} = P_{I2}S_{I1}$] $\frac{P_{I1}}{P_{G1} \cdot (P_{I1}S_{I2})} < \frac{P_{I2}}{P_{G2} \cdot (P_{I2}S_{I1})}$

(5) Simplify and rearrange terms. $P_{G2}S_{I1} < P_{G1}S_{I2}$

(6) Add equivalent terms. [From (c), we know $P_{G2}S_{G1} = P_{G1}S_{G2}$} $P_{G2}S_{I1} + P_{G2}S_{G1} < P_{G1}S_{I2} + P_{G1}S_{G2}$

(7) Simplify. $P_{G2}(S_{I1} + S_{G1}) < P_{G1}(S_{I2} + S_{G2})$

(8) Apply Stops definitions. $P_{G2}S_1 < P_{G1}S_2$

(9) Rearrange. $\frac{S_1/S_2}{P_{G1}/P_{G2}} < 1 \quad Q.E.D\ (1)$

Objective 2. Prove: $\frac{S_1/S}{P_{G1}/P_G} < 1$

(10) Continuing from step (8), add identical terms. $P_{G2}S_1 + P_{G1}S_1 < P_{G1}S_2 + P_{G1}S_1$

(11) Simplify. $S_1(P_{G2} + P_{G1}) < P_{G1}(S_2 + S_1)$

(12) Apply Stops & Population definitions. $S_1P_G < P_{G1}S$

(13) Rearrange. $\frac{S_1/S}{P_{G1}/P_G} < 1 \quad Q.E.D\ (2)$

Proof 2

This proof demonstrates that if the Stop Ratio/Crime Ratio measure is less than 1, the Stop Share/Crime Share measure is greater than the Ratio/Ratio measure. Above, in the second half of Proof 1, it was demonstrated that under this condition the Share/Share measure is also less than 1. Both proofs also work if all inequalities are reversed. Taken together, the implication is that the Stop Share/Crime Share measure is always closer to 1 than the Stop Ratio/Crime Ratio is, unless both are exactly 1. Definitions are the same as in Proof 1.

Assumption: $\frac{S_1/S_2}{P_{G1}/P_{G2}} < 1$	Prove: $\frac{S_1/S_2}{P_{G1}/P_{G2}} < \frac{S_1/(S_1+S_2)}{P_{G1}/(P_{G1}+P_{G2})}$

(1) Rearrange terms. $S_1P_{G2} < S_2P_{G1}$

(2) Add identical terms. $S_1P_{G2} + S_2P_{G2} < S_2P_{G1} + S_2P_{G2}$

(3) Simplify. $P_{G2}(S_1+S_2) < S_2(P_{G1}+P_{G2})$

(4) Rearrange terms. $\frac{P_{G2}}{S_2} < \frac{P_{G1}+P_{G2}}{S_1+S_2}$

(5) Multiply both sides by $\frac{S_1}{P_{G1}}$ $\frac{S_1P_{G2}}{S_2P_{G1}} < \frac{S_1(P_{G1}+P_{G2})}{P_{G1}(S_1+S_2)}$

(6) Rearrange terms. $\frac{S_1/S_2}{P_{G1}/P_{G2}} < \frac{S_1/(S_1+S_2)}{P_{G1}/(P_{G1}+P_{G2})}$ $Q.E.D.$

Proof 3a.

Definitions are the same as in Proof 1. The first step in this proof shows that the Stop Ratio/Crime Ratio measure is always less than the Average Stop Rate Ratio as defined in the notes to Table A2 above, with the higher-crime group (labeled Group 1 here) in the numerator of all ratios. Recall that this version of the Average Stop Rate Ratio represents the average effect of the inter-group conditional stop probability differences on Group 1. (Using the parlance of the reweighting literature, this could be called the "average treatment effect on the treated," expressed as a likelihood ratio.) The numerator of the ratio is the observed number of stops in Group 1; the denominator is the number that would have been observed in Group 1 if the distribution of guilt and innocence had been the same, but Group 2's stop probabilities (conditional on criminal conduct) had applied instead.

Assumptions

(a)	Group 1 has a higher crime rate, so: $\frac{P_{G1}}{P_{G2}} > \frac{P_{I1}}{P_{I2}}$
(b)	For both groups, P_I, P_G, S_I, and S_G are > 0.

$$\text{Prove: } \frac{S_1/S_2}{P_{G1}/P_{G2}} < \frac{S_1}{P_{I1} * S_{I2}/P_{I2} + P_{G1} * S_{G2}/P_{G2}}$$

(1) Rearrange terms from (a). $$P_{G1} > \frac{P_{G2}P_{I1}}{P_{I2}}$$

(2) Multiply by identical terms. $$P_{G1}S_{I2} > \frac{P_{G2}P_{I1}S_{I2}}{P_{I2}}$$

(3) Apply Stops definition. $$P_{G1}(S_2 - S_{G2}) > \frac{P_{G2}P_{I1}S_{I2}}{P_{I2}}$$

(4) Rearrange. $$P_{G1}S_2 > \frac{P_{G2}P_{I1}S_{I2}}{P_{I2}} + P_{G1}S_{G2}$$

(5) Divide both sides by $P_{G2}S_1$. $$\frac{P_{G1}S_2}{P_{G2}S_1} > \frac{P_{I1} * S_{I2}/P_{I2} + P_{G1} * S_{G2}/P_{G2}}{S_1}$$

(6) Invert both sides and reverse inequality. $$\frac{S_1/S_2}{P_{G1}/P_{G2}} < \frac{S_1}{P_{I1} * S_{I2}/P_{I2} + P_{G1} * S_{G2}/P_{G2}} \quad Q.E.D.$$

Proof 3b.

This next proof, which proceeds along extremely similar lines, shows that the Stop Ratio/Crime Ratio measure is also always less than an alternative measure of average stop-rate disproportionality conditional on criminal conduct—one that holds guilt rates constant at Group 2's average instead. This measure represents the likelihood ratio associated with applying Group 1's conditional stop probabilities (instead of Group 2's)

to Group 2 (which could be called the "average treatment effect on the untreated"). Assumptions and definitions are the same as above. Note that the overall "average treatment effect" for the population as a whole would be an average of these two versions of the Average Stop Rate Ratio, weighted by the sizes of Groups 1 and 2. It thus follows from Proofs 3a and 3b that that overall average would also always be higher than the Stop Ratio/Crime Ratio measure.

$$\text{Prove: } \frac{S_1/S_2}{P_{G1}/P_{G2}} < \frac{P_{I2} * S_{I1}/P_{I1} + P_{G2} * S_{G1}/P_{G1}}{S_2}$$

(1) Rearrange terms from (a). $$P_{G2} < \frac{P_{G1}P_{I2}}{P_{I1}}$$

(2) Multiply by identical terms. $$P_{G2}S_{I1} < \frac{P_{G1}P_{I2}S_{I1}}{P_{I1}}$$

(3) Apply Stops definition. $$P_{G2}(S_1 - S_{G1}) < \frac{P_{G1}P_{I2}S_{I1}}{P_{I1}}$$

(5) Rearrange. $$P_{G2}S_1 < \frac{P_{G1}P_{I2}S_{I1}}{P_{I1}} + P_{G2}S_{G1}$$

(6) Divide both sides by $P_{G1}S_2$ and rearrange. $$\frac{S_1/S_2}{P_{G1}/P_{G2}} < \frac{P_{I2} * S_{I1}/P_{I1} + P_{G2} * S_{G1}/P_{G1}}{S_2} \; Q.E.D.$$