University of Chicago Legal Forum

Volume 2016 Article 12

2016

# Testing Racial Profiling: Empirical Assessment of Disparate Treatment by Police

Sonja B. Starr

Follow this and additional works at: http://chicagounbound.uchicago.edu/uclf

Recommended Citation

Starr, Sonja B. (2016) "Testing Racial Profiling: Empirical Assessment of Disparate Treatment by Police," *University of Chicago Legal Forum*: Vol. 2016 , Article 12.
Available at: http://chicagounbound.uchicago.edu/uclf/vol2016/iss1/12

This Article is brought to you for free and open access by Chicago Unbound. It has been accepted for inclusion in University of Chicago Legal Forum by an authorized editor of Chicago Unbound. For more information, please contact unbound@law.uchicago.edu.

# Testing Racial Profiling: Empirical Assessment of Disparate Treatment by Police

*Sonja B. Starr*†

Statistical evidence plays a central role in litigation, scholarship, and public debates about race and policing. At one level, the statistical picture is clear: people of color in the United States, especially black men, interact with police far more often than white Americans do. Black Americans are about 2.5 times more likely to be arrested each year as their white counterparts.[1] Local studies show even larger racial disparities in the frequency of stops and use of force, although there are no national numbers.[2]

But while these gaps' existence is not contested, the reasons for them are. An especially hotly disputed question is whether and to what

† Professor of Law, University of Michigan. For comments and helpful discussions on prior versions, I am grateful to Alicia Davis, Avlana Eisenberg, Mark Fancher, Jim Greiner, Sam Gross, Louis Kaplow, Randy Kennedy, Anup Malani, Jonathan Masur, David Moran, J.J. Prescott, Eve Brensike Primus, Jon Sacks, Margo Schlanger, Michael Steinberg, Matthew Stephenson, and Kim Thomas, as well as *Legal Forum* participants and workshop participants at Harvard, University of Chicago, University of Michigan, University of Texas, University of Wisconsin, University of Colorado, and UC-Berkeley. Brian Apel, Grady Bridges, Alex Harris, Avi Kupfer, Linfeng Li, and Andrew Sand provided excellent research assistance.

[1] According to the Bureau of Justice Statistics Arrest Data Analysis Tool, the 2012 arrest rate was ten percent for black adults and four percent for whites; disparities are larger for more serious crimes. The Bureau of Justice Statistics's national estimates are not broken down by Hispanic ethnicity, or by race and sex combined. *Arrest Data Analysis Tool*, BUREAU OF JUST. STATISTICS, http://www.bjs.gov/index.cfm?ty=datool&surl=/arrests/index.cfm# (follow "National Estimates" hyperlink, then follow "Trend Tables by Race" hyperlink).

[2] *See, e.g.*, ACLU, BLACK, BROWN, AND TARGETED: A REPORT ON BOSTON POLICE DEPARTMENT STREET ENCOUNTERS FROM 2007–2010, 1 (2014), https://aclum.org/app/uploads/2015/06/reports-black-brown-and-targeted.pdf [https://perma.cc/GT9J-QTSC] (finding large disparities in pedestrian stop-and-frisk rates in Boston); *see* Jodi M. Brown & Patrick A. Langan, BUREAU OF JUST. STATISTICS, NCJ 180987, POLICING AND HOMICIDE, 1976-1998: JUSTIFIABLE HOMICIDE BY POLICE, POLICE OFFICERS MURDERED BY FELONS (2001), at iii (finding blacks four times as likely as whites to be killed by police); Bernard E. Harcourt, *Rethinking Racial Profiling: A Critique of the Economics, Civil Liberties, and Constitutional Literature, and of Criminal Profiling More Generally*, 71 U. CHI. L. REV. 1275, 1275–76 (2004) (describing traffic stop disparities); Rachel Harmon, *Why Do We (Still) Lack Data on Policing?*, 96 MARQ. L. REV. 1119, 1139–40 (2013) (calling for better data collection). A new, federally funded initiative seeks to build a national database on stops and use of force. Ctr. for Policing Equity, *Nation's First Police Profiling Database Awarded Grant By NSF* (Nov. 7, 2013), http://policingequity.org/wp-content/uploads/2013/12/database_release_final.pdf [https://perma.cc/XW8B-ZJLK].

extent these disparities result from police discrimination on the basis of race. That question, which is the central constitutional issue under longstanding equal protection doctrine, sharply divides public opinion—largely along racial lines.[3] Among commentators, polar opposite answers are each often presented as indisputable.[4] In part, these conflicts persist because the question is very challenging to answer empirically, due to data limitations and challenges of causal inference.

In this Article, I explore why measuring disparate-treatment discrimination by police is so difficult, and consider the ways that researchers' existing tools can make headway on these challenges and the ways they fall short. Lab experiments have provided useful information about implicit racial bias, but they cannot directly tell us how these biases actually affect real-world behavior. Meanwhile, for observational researchers, there are various hurdles, but the hardest one to overcome is generally the absence of data on the citizen conduct that at least partially shapes policing decisions. Most crime, and certainly most noncriminal "suspicious" or probable-cause-generating behavior, goes unreported and undetected, and is unobservable to researchers. The available measures of *observed* crime are not necessarily good proxies for total crime, and in any event, such data generally do not exist at an individual level that can be linked to individual outcome data on police interactions. Meanwhile, while we often *do* have data on the subset of people who are stopped by police, analyses limited to those individuals are often distorted by selection bias and by the absence of exogenous measures of their conduct; researchers have no choice but to rely, circularly, on what police write down.

These hurdles are serious. Some headway has been made in particular contexts in which quasi-experimental methods or direct physical observation by researchers is possible, but most policing contexts are not readily amenable to these approaches. It may be possible to do more using survey methods, though these pose their own

---

[3] *E.g.*, Ronald Weitzer & Steven A. Tuch, *Racially Biased Policing: Determinants of Citizen Perceptions*, 83 SOC. F. 1009, 1017 (2005) (finding blacks six times likelier than whites to believe police prejudice is a problem in their city).

[4] For example, a recent letter from civil rights and community leaders called the pattern of "aggressive police tactics [against young black men] . . . too obvious to be a coincidence. . . . [I]t is time for the country to counter the effects of systemic racial bias." Letter from Maya Rockeymoore et al., to President Obama (Aug. 25, 2014), http://www.washingtonpost.com/wp-srv/ad/public/static/letter/ [https://perma.cc/95ZD-BTBZ]. By contrast, prominent commentator Heather McDonald 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*, NBC (Aug. 17, 2014), http://www.nbcnews.com/meet-the-press/meet-press-transcript-august-17-2014-n182641 [https://perma.cc/36HS-HFTR].

challenges. And when it comes to assessing discrimination against *neighborhoods* of color (as opposed to individuals), it is sometimes possible to rely on aggregate-level data to make plausible claims. Often, however, the limits of available data will mean that it is just not possible to determine whether the police are discriminating based on race. These research challenges are also problems for courts, litigants challenging such discrimination, and police departments themselves as they seek to comply with their constitutional obligations.

I suggest, in some contexts, that a new approach would work better. The method I propose is called "auditing," which would employ "testers" (probably undercover officers) of different races to elicit possible interactions with the police. Auditing has not been tried or even discussed in the law enforcement field, which is surprising because for decades it has been a central tool in antidiscrimination research and civil rights enforcement more generally. It presents safety, legality, and efficacy concerns when applied to policing, but with careful design I argue that these concerns can be overcome. If so, auditing could provide something observational research usually cannot: causally rigorous analysis of police discrimination in a real-world setting.

Part I begins by examining why it is important to develop good methods for measuring "disparate treatment" discrimination by police. Disparate treatment is certainly not the *only* source of racial disparity in policing that researchers or policymakers should care about. That said, constitutional doctrine forces us to confront the question, and I outline other moral and policy reasons for why we should be concerned about disparate treatment. I also examine the conceptual problems associated with thinking of racial discrimination as a "cause" of disparity. In Part II, I examine existing methods of analyzing disparate treatment: individual- and neighborhood-level regression analyses, quasi-experimental methods exploiting variation in police ability to observe race, and lab experiments on implicit bias. In Part III, I set forth the auditing proposal and explore its advantages, challenges, and limitations.

## I. WHY MEASURING DISPARATE TREATMENT MATTERS

This paper addresses methods of estimating something quite specific: police racial discrimination of the "disparate treatment" variety, in the sense that U.S. courts use that term.[5] By this, I mean

[5] I use the terms "disparate treatment" and "discrimination" interchangeably throughout much of the paper, although "discrimination" can also entail broader meanings.

the extent to which police treat persons who are otherwise similarly situated (along relevant dimensions that the police perceive) differently because of race. Disparate treatment by police includes what is commonly called racial profiling: that is, disparate treatment that is based on race-based assumptions about differential crime risk. It also could encompass any other way in which racial perceptions consciously or unconsciously affect the decision-making of police vis-à-vis individuals or communities. In Sections A and B, respectively, I briefly outline some legal and policy reasons that disparate treatment discrimination is an important target of empirical estimation. In Section C, I make clear that I do not think this is the *only* valid way of conceptualizing racial inequality in policing, and I distinguish it from other conceptions that are also worthy subjects of empirical, legal, and policy analysis. Finally, claims of disparate treatment are causal in nature, and in Section D, I unpack what it means to treat race as a "cause" in this way.

### A. Racial Profiling and Constitutional Doctrine

Why do we need good empirical estimates of racially disparate treatment by police officers? The most obvious reason is that the existence of governmental disparate treatment is the central question posed by current equal protection doctrine. Current doctrine precludes constitutional challenges solely premised on racially disparate *impact* (i.e., differential effects on different racial groups)[6] or discrimination by private actors like witnesses.[7] But, as I show here, police racial discrimination essentially always violates the Equal Protection Clause. It is, however, difficult to prove, which makes effective empirical strategies especially important.

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.[8] These critics have grounds for

---

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

[7] One could argue that when the police give effect to private discrimination by carrying out stops and arrests, state action is generated. *Cf.* Shelley v. Kraemer, 334 U.S. 1, 19–21 (1948) (barring judicial enforcement of 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 extension of *Shelley*), and would also set a difficult standard for police, who may not know when witnesses are racially biased.

[8] *E.g.*, Delores Jones-Brown & Brian A. Maule, *Racially Biased Policing*, *in* RACE, ETHNICITY, AND POLICING 140, 141–43 (2010) (Stephen K. Rice & Michael D. White eds. 2010); Albert W. Alschuler, *Racial Profiling and the Constitution*, 2002 U. CHI. LEGAL F. 163, 164–66; Angela J. Davis, *Race, Cops, and Traffic Stops*, 51 U. MIAMI L. REV. 425, 442–43 (1997); Evan

frustration: the Court has avoided squarely deciding whether the Equal Protection Clause bars racial profiling 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.[9] Still, as I show here, 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 in other cases outside the policing context.[10] It is even more obvious that it would be unconstitutional for police to discriminate on the basis of some other type of racial bias *unrelated* to crime prevention aims, so I will focus on the racial profiling issue here.

Scholars examining the relevant constitutional doctrine have mainly focused on the Court's numerous adverse Fourth Amendment precedents.[11] These include *Whren v. United States*,[12] which held that a traffic stop provides probable cause for a vehicle search even if the traffic violation was a mere pretext, and *United States v. Brignoni-Ponce*,[13] which suggested that Mexican appearance might provide reasonable suspicion of an immigration violation when combined with other factors, but not alone. These cases were indeed big setbacks for those challenging racial profiling, but they do not directly implicate equal protection claims. *Brignoni-Ponce* sent a confusing signal (why suggest that ethnicity may be relevant to Fourth Amendment analysis

---

Gerstman & Christopher Shortell, *The Many Faces of Strict Scrutiny*, 72 U. PITT. L. REV. 1, 46–50 (2001); Kevin R. Johnson, *How Racial Profiling in America Became the Law of the Land:* United States v. Brignoni-Ponce *and* Whren v. United States *and the Need for Truly Rebellious Lawyering*, 98 GEO. L.J. 1005, 1006 (2010); Meaghan Paulhamus et al., *State of the Science in Racial Profiling Research*, *in* RACE, ETHNICITY, AND POLICING 239, 242–43.

[9] Notoriously, in Brown v. City of Oneonta, 221 F.3d 329 (2d Cir. 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 8, at 47; Alschuler, *supra* note 8, at 179–92. However, courts do consistently distinguish the use of race in specific suspect descriptions from using race to make behavioral generalizations about broad 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).

[10] The Sixth Circuit has gotten this issue wrong, however. *E.g.*, United States v. Travis, 62 F.3d 170, 174 (6th 1995); *see* Alschuler, *supra* note 8, at 178–79 (critiquing this and other cases).

[11] *E.g.*, Jones-Brown & Maule, *supra* note 8, at 140–57; Jeffrey A. Fagan et al., *Street Stops and Broken Windows Revisited*, *in* RACE, ETHNICITY, AND POLICING, *supra* note 8, at 309, 312–13; Johnson, *supra* note 8, at 1006–08.

[12] 517 U.S. 806, 813–16 (1996).

[13] 422 U.S. 873, 885–87 (1975); *see* United States v. Martinez–Fuerte, 428 U.S. 543, 563 (1976) (holding that the authority of police to consider Mexican ancestry is greater during stops made at checkpoints than during roving stops).

if its consideration is barred by the Fourteenth?), but that possible signal does not trump the more directly relevant equal protection precedents striking down decision-makers' use of race even when it is combined with other factors.[14] Many scholars have critiqued the doctrinal separation of Fourth and Fourteenth Amendment objections to racial profiling,[15] but the upside of this approach is that adverse Fourth Amendment holdings do not decide the Fourteenth Amendment issue.[16]

The Supreme Court has never directly decided whether racial profiling by the police violates the Equal Protection Clause, but to say that it does not, it would have to upend decades of doctrine. The key line of cases has arisen in contexts outside policing, but its principles are directly applicable. It concerns the prohibition of "statistical discrimination." When applying heightened scrutiny, 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.[17]

For example, in *Craig v. Boren*,[18] the Court struck down a law establishing different minimum drinking ages for 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. Similarly, the Court has struck down governmental reliance on gendered or racial generalizations about learning styles,[19] juror voting,[20] and workforce participation.[21] All these

---

[14] *E.g.*, Arlington Heights v. Metropolitan Housing Dev. Corp., 429 U.S. 252, 265–66 (1977).

[15] *E.g.*, Alschuler, *supra* note 8, 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).

[16] Scholars have suggested that *Whren* and related cases green-light racial profiling in car searches. *See* Jones-Brown & Maule, *supra* note 8, 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)); Fagan et al., *supra* note 11, at 312; Paulhamus et al., *supra* note 8, at 242–43. This is likely often true in practice, because it was a huge blow to Fourth Amendment claims. But while declining in the Fourth Amendment context to dig into police's true motives, *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").

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

[18] 429 U.S. 190, 202–04 (1976).

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

[20] 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).

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

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 within the group. The Court has carved out exceptions only for physical sex differences relating to pregnancy.[22] It has never made exceptions for generalizations about behavior, and it would be shocking if it did so for racial generalizations about criminal tendencies.

This line of cases should be fatal to the most likely constitutional defense of racial profiling, namely the claim that profiles are empirically supported and thus facilitate the police objective of preventing crime. This is so even assuming crime prevention is a compelling state interest.[23] 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.

In any event, law enforcement bodies have generally agreed that racial profiling is illegal. For example, in 2003, the U.S. Department of Justice declared it "absolutely prohibited."[24] The remaining ambiguity in the case law may therefore be irrelevant in practice. Modern police departments do not defend racial profiling. They deny that they engage

---

[22] *See, e.g.*, Tuan Anh Nguyen v. I.N.S., 533 U.S. 53, 68 (2001). Note that, while race-based classifications are subject to strict scrutiny, gender-based classifications are subject to an intermediate standard of review due to "enduring" physical differences between men and women. *Virginia*, 518 U.S at 533.

[23] *See Virginia*, 518 U.S. at 531 (listing prohibition on gender generalizations and a substantial relationship to important government interests as separate requirements).

[24] 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 [https://perma.cc/WQR6-GHJ3]. A 2014 guidance qualifies this prohibition under limited circumstances for national-security-related screenings, in particular when looking for persons suspected to be associated with a particular terrorist or criminal organization "whose membership has been identified as overwhelmingly" being of a particular race. U.S. DEP'T OF JUSTICE, GUIDANCE FOR FEDERAL LAW ENFORCEMENT AGENCIES REGARDING THE USE OF RACE, ETHNICITY, GENDER, NATIONAL ORIGIN, RELIGION, SEXUAL ORIENTATION, OR GENDER IDENTITY 2 (2014), https://www.justice.gov/sites/default/files/ag/pages/attachments/2014/12/08/use-of-race-policy.pdf [https://perma.cc/R8RC-ELQU]. However, the guidance explains that even in this context officers must avoid "invidious profiling" and must instead rely on specific "trustworthy information, relevant to the locality or time frame"—for example, reliable information that members of a "foreign ethnic insurgent group" are planning a suicide bombing targeting the president of that foreign country during a state visit. *Id.* at 9–10.

in it.[25] Settlements in racial profiling lawsuits frequently contain terms prohibiting "any consideration of race."[26]

If the Fourteenth Amendment argument is doctrinally well-supported and not in practice contested, why do litigants routinely not win Fourteenth Amendment challenges? And why has the Fourth Amendment played a more prominent role in profiling litigation? The key reason is evidentiary: it is very hard for litigants to prove racial profiling.[27] Individual criminal defendants raising selective-enforcement defenses face especially steep hurdles.[28] In federal criminal cases, just getting discovery is notoriously difficult.[29] Even if defendants can show a broad pattern of discrimination, they must also show that it affected their cases specifically. Statistical evidence almost

---

[25] *E.g.*, Melendres v. Arpaio, 695 F.3d 990, 995 (9th Cir. 2012) (describing Arizona sheriff's defense to equal protection suit: "[d]efendants do not engage in racial profiling"); *PBS NewsHour*, (PBS broadcast Aug. 13, 2013), http://www.pbs.org/newshour/bb/nation-july-dec13-stopfrisk_08-13/ [https://perma.cc/L4ZS-CVVS] (quoting NYPD Commissioner Raymond Kelly: "We do not engage in racial profiling. It is prohibited by law [and] by our own regulations."); Greg Risling, Associated Press, *DOJ Finds 2 LA Sheriff's Stations Discriminating*, SAN DIEGO TRIB. (June 28, 2013), http://www.sandiegouniontribune.com/news/2013/jun/28/doj-finds-2-la-sheriffs-stations-discriminating/ [https://perma.cc/VP7R-K8CC] (quoting L.A. Sheriff Department spokeswoman Steve Whitmore: "We stand resolute that we have not discriminated against members of the public[.]"); Jane Prendergast, *Officers' Hearts Hold Racial Profiling Solution, Chief Says*, CIN. ENQUIRER (Mar. 6, 2001), http://enquirer.com/editions/2001/03/06/loc_officers_hearts_hold.html [https://perma.cc/8QZG-WXUD] (quoting Cincinnati police chief: Profiling "is not only wrong, it's unconstitutional. It's illegal. We know that. We teach that."); Sho Wills, *Chicago, New York Officers Spar Over Stop-and-Frisk Policy*, CNN (Aug. 14, 2014), http://www.cnn.com/2013/08/14/us/new-york-chicago-stop-frisk/ [https://perma.cc/FDW6-4933] (quoting Chicago Police Department spokesman Adam Collins: "[W]e don't engage in racial profiling."); Letter from S.C. Kitchen, Defense Attorney, to Assistant U.S. Attorney General Thomas E. Perez (Sept. 26, 2013), http://www.timesnewshosting.com/docs/johnson.pdf [https://perma.cc/P8FF-49YF] (denying allegations that Terry Johnson, Sheriff of Alamance County, had engaged in racial profiling).

[26] *See* Sam R. Gross & Katherine Y. Barnes, *Road Work: Profiling and Drug Interdiction on the Highway*, 101 MICH. L. REV. 651, 741–43 (2002).

[27] *See id.* at 653–57, 741; Johnson, *supra* note 8, at 1063–64; David Rudovsky, *Law Enforcement by Stereotypes and Serendipity: Racial Profiling and Stops and Searches without Cause*, 3 U. PA. J. CONST. L. 296, 322–29 (2001); Sklansky, *supra* note 15, at 326 ("[C]hallenges to discriminatory police practices will fail without proof of conscious racial animus on the part of the police . . . . [T]his amounts to saying that they will almost always fail.").

[28] *See, e.g.*, United States v. Armstrong, 517 U.S. 456, 465 (1996) (Supreme Court denied a motion for discovery on a selective prosecution claim because plaintiff failed to show that "similarly situated individuals of a different race were not prosecuted."); *see also* RANDALL KENNEDY, RACE, CRIME, AND THE LAW 354 (1997) ("Research has uncovered no cases" of convictions overturned for selective prosecution, as of that date.).

[29] The Supreme Court has required "some evidence" of "differential treatment of similarly situated members of other races." *Armstrong*, 517 U.S. at 465–67. *Armstrong* addressed a claim of selective *prosecution*, and the Supreme Court has never specifically held that it applies to disparate-policing cases; in my view, it should not. Identifying a "similarly situated" group is likely especially difficult in policing cases: police keep no "records of instances in which they could have stopped a motorist . . . but did not." Davis, *supra* note 8, at 438.

never clears this hurdle alone, though it might help in combination with case-specific qualitative evidence.[30]

For these reasons, the best prospects for equal protection challenges to succeed are in civil cases (class actions or government enforcement actions), in which the pattern of discrimination is the issue. Such cases can and have succeeded, and have had important consequences for police practices; the *Floyd v. City of New York*[31] litigation, which helped to bring about the New York Police Department's (NYPD) massive reduction in stop-and-frisk practices, is a recent example.[32] Such claims turn centrally on statistical evidence.

## B. Other Policy Reasons to Measure Police Disparate Treatment

Some commentators, while acknowledging the legal importance of the disparate-treatment question, have dismissed its moral importance. David Thacher, for example, describes the focus on "racial profiling" as a parochial concern of lawyers—a distraction from "substantive" equality.[33] Other critics have dismissed the focus on intentional discrimination as "legalistic."[34] Moreover, beyond the policing context,

---

[30] In principle, strong statistical evidence could allow an inference that the defendant *probably* would not have been stopped but for race; this would be the logical inference if a defendant's race made him more than twice as likely to be stopped. But courts have resisted this sort of reasoning, *see* Harcourt, *supra* note 2, 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; in some other criminal-law contexts where equal protection claims were made, the Court has been more receptive to statistical evidence of discrimination. *E.g.*, Castaneda v. Partida, 430 U.S. 482 (1977); Whitus v. Georgia, 385 U.S. 545 (1967).

[31] 959 F. Supp. 2d 540 (S.D.N.Y. 2013).

[32] It is hard to disentangle the relative influence of the *Floyd* lawsuit from that of broader changes in politics and police department policy, especially because the highly public lawsuit and the evidence of disparities that plaintiffs provided may have shaped the politics of the stop-and-frisk debate. After Mayor DeBlasio's election, the City dropped its appeal and agreed to a three-year monitoring plan in early 2014. Ray Sanchez et al., *New York Drops Appeal of Controversial Stop-and-Frisk Ruling*, CNN (Jan. 30, 2014), http://www.cnn.com/2014/01/30/us/new-york-drops-stop-frisk-appeal/ [https://perma.cc/H3AR-VBVA]. Stop-and-frisk rates, already declining during the period the lawsuit was pending, continued to decline sharply, and in 2015 they were an estimated ninety-six percent below their peak in 2011. *See Stop-and-Frisk Down, Safety Up*, NYCLU (Dec. 10, 2015), http://www.nyclu.org/news/stop-and-frisk-down-safety-nyclu-data-analysis [https://perma.cc/QGR8-GGPZ].

[33] David Thacher, From Racial Profiling to Racial Equality 1–2 (Aug. 2002) (unpublished manuscript) http://fordschool.umich.edu/research/papers/PDFfiles/02-006.pdf [https://perma.cc/X2CF-49KK].

[34] Robin S. Engel, *A Critique of the 'Outcome Test' in Racial Profiling Research*, 25 JUST. Q. 1, 5–9 (2008). This and some similar critiques, however, seem to caricature the "legalistic"

legal scholars have also long critiqued the focus on colorblindness, arguing that equality law should primarily target group subordination, not forbidden classifications.[35] On this view, use of racial classifications may be appropriate if they are invoked for a purpose that helps to promote substantive equality—the proper objective is not to be *blind* to race, but rather to acknowledge and seek to reduce racial stratification. From this perspective, a focus on disparate treatment discrimination can be critiqued for the narrowness of its inquiry, and for its embrace of a "colorblindness" objective that—in other contexts—has been an obstacle to race-conscious efforts to promote a more substantive vision of racial equality.

I generally sympathize with this anti-subordination view. But racially disparate treatment by police is still worth worrying about and measuring, and not just for the practical reason that current equal protection doctrine demands it. Whether the police racially discriminate is not "merely" a legalistic concern. Racially disparate treatment adds a substantively meaningful dimension of harm (exacerbating substantive racial inequality), as well as a distinct target for policy interventions.

Critics of racial disparities in policing have emphasized the role of discrimination, "racial profiling," or just "racism."[36] This framing ought not to be dismissed as mere legalism; it is not only coming from lawyers and is not usually centrally motivated by legal doctrine. Rather, it has

---

perspective, confusing the normative claim that disparate treatment matters (and is worth measuring) with the empirical claim that such disparate treatment is necessarily at the root of all observed disparities. For example, Pickerill et al. seem to suggest 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. Engel, *supra*, at 7. 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 28, at 149–51.

[35] *E.g.*, Jack M. Balkin & Reva B. Siegel, *The American Civil Rights Tradition: Anticlassification or Antisubordination?*, 58 U. MIAMI L. REV. 9, 9 (2003) ("Antisubordination theorists contend that guarantees of equal citizenship cannot be realized under conditions of pervasive social stratification and argue that law should reform institutions and practices that enforce the secondary social status of historically oppressed groups.").

[36] *E.g.*, *The Targeting of Young Blacks by Law Enforcement: Ben Jealous in Conversation with Jamelle Bouie*, AM. PROSPECT (Fall 2014), http://prospect.org/article/targeting-young-blacks-law-enforcement-ben-jealous-conversation-jamelle-bouie [https://perma.cc/64C8-LYRH]; Rockeymoore et al., *supra* note 4; Press Release, Rep. John Lewis on Shooting of Michael Brown and Events in Ferguson, Missouri (Aug. 14, 2014), http://johnlewis.house.gov/press-release/rep-john-lewis-shooting-michael-brown-and-events-ferguson-missouri [https://perma.cc/NC9S-W8MB].

cultural resonance and moral force.[37] The harms of racially disparate policing are thus often substantially amplified by a sense of racial targeting. Perceptions of police racism also deeply undercut trust in the police in black communities, which may undermine police effectiveness.[38] In short, racially disparate treatment may be just one morally troubling cause of racial disparity, but it is an important one.[39]

The particular harms associated with racially disparate treatment are, if anything, amplified by the purported justification that implicitly or explicitly underlies it, namely generalizations about racial groups' crime rates. For the government to generalize that people of color are dangerous, and to specifically target them on that basis for surveillance and arrest, is expressively and morally noxious, especially because such generalizations have a painful history in our culture.[40] I am not suggesting that one should not acknowledge racial differences in crime rates where they exist. What is poisonous is using those differences to justify ignoring differences *within* groups, making law-abiding citizens "pay for fears generated by criminals with whom they are lumped by dint of color."[41] This expressive harm is part of the distinctive injury done by racial profiling specifically.

Moreover, it is practically useful to disentangle police disparate treatment from other causes of policing disparities, even if one views those other causes as normatively problematic as well. Teasing out different causes of disparity can help guide policy responses. For example, many police departments have recently invested effort in implicit-bias testing and training, an approach that assumes that policing disparity is at least partially grounded in the assumptions

---

[37] *See* REBECCA M. BLANK ET AL., MEASURING RACIAL DISCRIMINATION 103 (2004) (stating more generally that "the broader public vision of what discrimination means [is] the treatment of two (nearly) identical people differently").

[38] *See* KENNEDY, *supra* note 28, at 151–53; Tom R. Tyler & Jeffrey Fagan, *Legitimacy and Cooperation*, *in* RACE, ETHNICITY, AND POLICING, *supra* note 8, at 84, 102–04; Weitzer & Tuch, *supra* note 3, at 1017–18.

[39] To be sure, a "colorblind" policing objective could sometimes interfere with efforts to achieve other equality objectives, including equality conditional on criminal conduct. For example, if certain stop criteria are less predictive of guilt for one race than another, police might have to take race into account when interpreting them if they want to avoid imposing racially disparate impacts. I do not seek to resolve these dilemmas here, although my own leaning would be to permit race-conscious efforts to avoid disparate impacts, favoring a more substantive view of equality. The current Supreme Court might not agree, however. *See* Ricci v. DeStefano, 557 U.S. 557 (2009) (casting doubt on the constitutionality of race-conscious efforts to avoid disparate impact liability).

[40] *See, e.g.*, KENNEDY, *supra* note 28, at 16; Patricia Williams, *Spirit-Murdering the Messenger*, 42 U. MIAMI L. REV. 127, 129–30 (1987); President Barack Obama, Remarks by the President on Trayvon Martin (July 19, 2013), https://www.whitehouse.gov/the-press-office/2013/07/19/remarks-president-trayvon-martin [https://perma.cc/82B5-N7AC].

[41] KENNEDY, *supra* note 28, at xi.

officers make about people of color.[42] Empirical estimates of disparate-treatment discrimination could help to motivate this and other interventions that seek to alter police decision-making processes. On the other hand, to the extent that such analyses show that policing disparities are substantially grounded in other causes, they will help to warn us that interventions that merely seek "color-blind policing" may prove disappointingly unable to eliminate those disparities. In that case, alternative or additional policy interventions may be called for, and I discuss a couple of examples in the next Section.

### C. Distinguishing the Estimation of Disparate Treatment from Other Objectives of Policing-Disparity Research

Although estimation of disparate treatment by police is a worthwhile objective for empiricists, it is not the *only* worthwhile objective, and it bears emphasis that there are many normatively important sources of racial disparity in policing that this concept does not encompass. Several other conceptions of racial inequality in policing have been and should be the focus of empirical inquiry as well. For readers who are attempting to make sense of seemingly conflicting statistics concerning race and policing, it is important to clarify that not every study is attempting to measure the same thing. That is, different studies either implicitly or explicitly conceptualize racial inequality in different ways—some of which are quite different from disparate treatment discrimination. Here, I identify a few key examples.

#### 1. Raw racial disparity.

First, estimating disparate treatment is a narrower and much more difficult task than estimating raw racial *disparities* in police interactions. Raw disparity statistics entail simple comparisons across racial groups of the per capita rates of police interactions, or, similarly, comparisons of a group's population share to its share of police interactions. Such statistics have played an important role in debates about race and policing, including some empirical studies,[43] some legal scholarship,[44] and a large share of media coverage.[45] For example, one

---

[42] Tracey G. Gove, *Implicit Bias and Law Enforcement*, 78 POLICE CHIEF 44 (2011).

[43] *E.g.*, Jeff Rojek et al., *The Influence of Driver's Race on Traffic Stops in Missouri*, 7 POLICE Q. 126 (2004).

[44] *E.g.*, Bennet Cappers, *Rethinking the Fourth Amendment: Race, Citizenship, and the Equality Principle*, 46 HARV. C.R.-C.L. L. REV. 1, 14–19 (2011); Johnson, *supra* note 8; 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, 724 n.9 (2004).

recent study found: "Blacks were subjected to 63% of [pedestrian stops], even though they made up just 24% of Boston's population."[46]

Raw disparities are easy enough to measure, so long as police departments collect data about the police interaction in question and the demographics of those subjected to it. Such data are available on a national scale for arrests, but are only available in some jurisdictions for other interactions such as stops, searches, and use of force. This data shortcoming is a readily soluble problem, requiring only political will and commitment of some resources, and there is currently a significant effort funded by the U.S. Department of Justice to build a national database on stops and use of force.[47] Raw disparities are comparatively easy to estimate because, properly understood, they entail no causal claims and do not require measuring difficult-to-observe variables such as crime commission or suspicious behavior.

As critics often point out, identifying the existence of raw disparities does not tell us the *reasons* for them, and in particular does not prove police racial discrimination. But raw disparity statistics are very important for other purposes; most obviously, they demonstrate that people of color disproportionately bear the burdens of our criminal justice system and its expansion in recent decades. And while policing can obviously also bring benefits to communities, the burdens it imposes are substantial. Interacting with police is often stressful and scary,[48] and even if no charges are brought, arrest records can produce stigma, job-market consequences, and increased sentences in future cases.[49] If charges and punishment are pursued, the consequences of police interactions are obviously even greater, and much of the racial disparity in U.S. incarceration rates can be explained by disparate

---

[45] *E.g.*, Jess Bidgood, *Boston Police Focus on Blacks in Disproportionate Numbers, Study Shows*, N.Y. TIMES (Oct. 8, 2014), http://www.nytimes.com/2014/10/09/us/boston-police-focus-on-blacks-in-disproportionate-numbers-study-shows.html?_r=1 [https://perma.cc/7YVG-YYER]; *see* Greg Ridgeway & John MacDonald, *Methods for Assessing Racially Biased Policing*, *in* RACE, ETHNICITY, AND POLICING, *supra* note 8, at 180, 181 (describing the "compulsion in media reports" to focus on per-capita racial disparities in police stops).

[46] ACLU, *supra* note 2, at 1.

[47] *Nation's First Police Profiling Database Awarded Grant By NSF*, CTR. FOR POLICING EQUITY (Nov. 7, 2013), http://policingequity.org/wp-content/uploads/2013/12/database_release_final.pdf [https://perma.cc/XW8B-ZJLK].

[48] *See* Alschuler, *supra* note 8, at 212–13; Rod K. Brunson, *Beyond Stop Rates: Using Qualitative Methods to Examine Racially Biased Policing*, *in* RACE, ETHNICITY, AND POLICING, *supra* note 8, at 221, 224–33.

[49] *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), http://www.wsj.com/articles/as-arrest-records-rise-americans-find-consequences-can-last-a-lifetime-1408415402 [https://perma.cc/H2TE-69C5].

arrest rates.[50] These burdens are not merely borne by the guilty—especially the burdens of stops and searches, most of which produce no evidence of wrongdoing.[51]

Empirical research on raw disparities can add to our understanding of these burdens. And it provides an essential starting point for any further empirical assessment of *why* those disparities exist, and for a policy assessment of what can be done about them. Disparity research can motivate policy changes within the criminal justice system as well as broader changes outside of it, such as social policies addressing poverty and other root causes of crime-rate disparities.

2. Other criminal justice policies and practices that shape racial disparities.

Second, if researchers seek to go beyond simply measuring disparities, and identify causes of those disparities that can be addressed via changes to criminal justice policies, disparate treatment discrimination is not the only such cause that should be of interest. Rather, researchers can shed light on the racially disparate impacts of facially race-neutral policy choices, including the ways we define crimes, grade their severity, and apportion enforcement resources. Such studies have been relatively uncommon, but one example is Golub et al.'s study of NYPD's massive increase in marijuana enforcement during the 1990s; the authors show that this change very disproportionately affected black and Latino New Yorkers.[52] This disproportionate effect might or might not have been the product of police discrimination in the disparate treatment sense—Golub et al. did not seek to answer this question. But regardless of the answer, the choice to prioritize marijuana enforcement in the first place was a *choice*—one that did not have to be made, and could be reversed—which had strongly racially disparate consequences.

---

[50] *See, e.g.*, Brett E. Garland et al., *Racial Disproportionality in the American Prison Population*, 5 JUST. POL'Y J. 1, 14–25 (2008) (reviewing literature); Alfred Blumstein, *Racial Disproportionality of U.S. Prison Populations Revisited*, 64 U. COLO. L. REV. 743 (1993) (In 1991, seventy-six percent of the black-white incarceration gap stemmed from arrest.). Black incarceration rates are about six times the white rate. One in nine black men under age thrity-five is incarcerated. JENIFER WARREN ET. AL., PEW CTR. ON THE STATES, ONE IN 100: BEHIND BARS IN AMERICA 2008 3, 6 (Feb. 2008), http://www.pewtrusts.org/~/media/legacy/uploadedfiles/wwwpewtrustsorg/reports/sentencing_and_corrections/onein100pdf.pdf [https://perma.cc/A25F-KHLL].

[51] *See* David A. Harris, *The Stories, the Statistics, and the Law: Why "Driving While Black" Matters*, *in* RACE, ETHNICITY, AND POLICING, *supra* note 8, at 36, 49.

[52] Andrew Golub et al., *The Race/Ethnic Disparity in Misdemeanor Marijuana Arrests in New York City*, 6 CRIM. & PUB. POL'Y 131, 137 (2007).

3. Racial disparity unexplained by crime.

Often, efforts to explain policing disparities empirically assess only one explanatory variable other than race: crime commission. That is, researchers as well as lawyers, courts, and other commentators often ask: are racial disparities in stops, arrests, and other policing statistics explained by racial disparities in crime commission? These comparisons, which are made routinely by both defenders and critics of police departments, often take the form of comparing a racial group's share of crimes to its share of police interactions ("share comparisons"), with crime measured according to some benchmark such as police reports or survey data.[53] Alternatively, they sometimes take the form of comparing, across racial groups, the ratio of police interaction rates to crime rates ("ratio comparisons").[54]

In another paper, I offer an extensive critique of these sorts of comparisons, which often dramatically overstate crime's explanatory role by ignoring the fact that not all those who are subjected to police interactions are guilty.[55] If researchers want to assess whether criminal conduct explains policing disparity, "share comparisons" and "ratio comparisons" will be distorted by their failure to account for interactions with the innocent.[56] There are plausible alternative approaches, including regression analyses that include both race and some measure of criminal conduct as regressors. I consider challenges associated with this kind of analysis in Part II.A.

In any event, these simple three-variable comparisons (police interaction rates, race, crime), even if constructed in a more sensible way, should usually *not* be interpreted as estimates of police racial discrimination. The question of whether people with similar criminal

---

[53] *See, e.g.*, Robert J. Sampson & Janet L. Lauritsen, *Racial and Ethnic Disparities in Crime and Criminal Justice in the United States*, 21 CRIME & JUST. 311, 328 (1997) (comparing fifty-six percent black robbery suspect share to sixty-one percent arrest share); Stacey Patton, *If You're White, That Joint Probably Won't Lead to Jail Time*, WASH. POST (Jan. 12, 2014), https://www.washingtonpost.com/opinions/if-youre-white-that-joint-probably-wont-lead-to-jail-time/2014/01/10/caa94154-77f8-11e3-af7f-13bf0e9965f6_story.html [https://perma.cc/3TAE-N8PT] (comparing fourteen percent black drug-user share to thirty-four percent drug-arrest and fifty-three percent drug-incarceration shares).

[54] *See, e.g.*, BLANK ET AL., *supra* note 37, at 193 (comparing eighteen percent black speeding share to seventy-three percent search share).

[55] *See generally* Sonja Starr, *Race and Policing: How to Make Sense of the Numbers (And How Not to)*, Working Paper (on file with author), Part I; For additional commentary on this approach, see 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).

[56] *See generally* Sonja B. Starr, *Explaining Race Gaps in Policing: Normative and Empirical Challenges* Part II (Univ. of Mich., Working Paper No. 15-003, Jan. 19, 2015).

conduct are treated similarly by police may be morally important in its own right, and worthy of empirical estimation; I discuss the reasons this is so in my other paper.[57] But it is generally not the same as the disparate-treatment question, and it is not the focus of this piece. For the purpose of estimating racially disparate treatment, crime is one important potentially confounding variable, but not necessarily the only one. "Accounting for crime" alone might lead to either an overestimate or an underestimate of police racial discrimination, depending on which way other unobserved variables cut.

For this reason, in Part II.A, I assume that observational assessments of police discrimination will usually also seek to account for other variables. However, it *may* be reasonable to treat crime as the only relevant confounding variable in one (fairly common) situation: where the police themselves argue that policing disparities are explained by crime differences, and offer no additional explanations. Then, the analysis can be seen as testing whether the department's explanation holds up.

4. Irrational or "taste-based" discrimination.

While disparate-treatment discrimination is a relatively narrow way to conceptualize policing inequality, an even narrower conception underlies one common approach to the estimation of policing disparity. This approach compares across races the "hit rates" of police interactions—usually, the rate at which stops lead to arrests, which is taken as a proxy for guilt. Assuming police are motivated to maximize the number of arrests, equal hit rates across races are interpreted to show that police are considering race "rationally"; unequal hit rates imply irrational "taste-based" discrimination. Hit-rate models dominate the economic literature on policing disparities.[58] Elsewhere, I have critiqued these models on a number of fronts, arguing that they rely on faulty empirical assumptions and make wrong predictions.[59]

---

[57] *Id.*

[58] 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); Nicola Persico, *Racial Profiling, Fairness, and Effectiveness of Policing*, 92 AM. ECON. REV. 1472, 1479 (2002); *cf.* GARY S. BECKER, ACCOUNTING FOR TASTES 140–42 (1996) (providing the central insight that "tastes for discrimination" are generally not efficient to discriminators).

[59] Starr, *supra* note 55, at Part II. For additional criticism, see also Harcourt, *supra* note 2, at 1295–1314 (criticizing the economic models of racial profiling because (1) the models incorrectly define "success" as maximizing the total number of arrests, and (2) the models assume that the official criminal rates are a good proxy for real offending rates).

Here, however, I will simply observe that estimating "taste-based" discrimination is not the same as estimating racially disparate treatment. The theory underlying hit-rate models distinguishes taste-based discrimination from statistical discrimination (use of race as a proxy for a legitimate consideration with which it is correlated—here, criminality). Hit-rate models assume that racially unbiased police *will* consider race when assessing how suspicious somebody is—that is, the likelihood of criminality—if indeed there are racial differences in crime rates. Economists sometimes defend statistical discrimination as efficient, although the conditions under which it is so have been much debated.[60]

But as discussed in Section A, U.S. constitutional law draws no distinction between racially disparate treatment that is grounded in statistical generalizations and racially disparate treatment that is grounded in mere preference or prejudice. Both are components of unconstitutional discrimination, and so an approach that measures only the latter will be underinclusive, even assuming other empirical concerns can be set aside. For this reason, I do not include the hit-rate studies in Part II's review of methods of estimating police discrimination: they are unsuited to the task at hand.

### D. On Race and Causation

Estimates of disparate-treatment discrimination are estimates of *causal* effects, not mere correlations—specifically, the causal effect of citizens' race (or of the racial compositions of communities) on police decision-making.[61] But identifying causal effects of race is challenging in practice and perhaps conceptually as well, 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.[62] Perhaps we cannot sensibly ask how a person's

---

[60] *See* Peter Norman, *Statistical Discrimination and Efficiency*, 70 REV. ECON. STUD. 615 (2003); Edmund S. Phelps, *The Statistical Theory of Racism and Sexism*, 62 AM. ECON. REV. 659 (1972) (providing a seminal theoretical account); Stewart Schwab, *Is Statistical Discrimination Efficient?*, 76 AM. ECON. REV. 228 (1986).

[61] *See, e.g.*, Lincoln Quillian, *New Approaches to Understanding Racial Prejudice and Discrimination*, 32 ANN. REV. SOC. 299, 302 (2006) (defining "discrimination" as "the causal effect of race").

[62] D. James Greiner & Donald B. Rubin, *Causal Effects of Perceived Immutable Characteristics*, 93 REV. ECON. & STAT. 775, 783–84 (2011); Maya Sen & Omar Wasow, *Race as a Bundle of Sticks: Designs that Estimate Effects of Seemingly Immutable Characteristics*, 19 ANN. REV. POL. SCI. 499 (2016).

outcome would have differed but for her race if her entire life would have been different in that counterfactual.[63]

This conceptual hurdle is not insuperable, however. Usually, when we ask causal questions about racial discrimination, we are not asking about the lifelong effects of race, but rather about discrimination in a particular decision process (e.g., arrest). The counterfactual is how the decision-maker would have responded had she encountered a person of a different race whose relevant characteristics (as perceived by the officer) were otherwise similar. To assess this question, it does not really matter how or why the individual developed those observable characteristics, and in particular whether race affected the likelihood of his doing so. This counterfactual analysis need not entail imagining, for example, a "white version" of a black citizen, identical in all ways but race. It entails merely comparing the way police treat people with similar *observable and decision-relevant* characteristics—for example, white and black drivers who are driving the same speed on a given highway.

James Greiner and Donald Rubin have suggested referring to "perceived race" to highlight the fact that we are talking about the decision-maker's perspective.[64] Another way (slightly clearer, in my view) is to say that we are assessing the "effects of police racial discrimination," if police are the decision-makers of interest. In any event, though, I think there is not much harm in the shorthand "effect of race," provided we are clear on what it means. Note that if one *did* want to examine "the effect of dynamic, cumulative discrimination" throughout a person's life (and, even further back, the life of his family and community), the strategies discussed in the next Part would not much help, but raw disparity estimates might.[65]

Moreover, in my view, it is not only conceptually possible to assess the effects of racial perceptions on police treatment of individual communities; it is important to do so, for reasons outlined in Sections A and B above. Of course, we should think about race as "causing" policing disparities through multiple channels, of which police racial discrimination is just one. When we estimate disparate treatment discrimination by police, we should remember that components of racial

---

[63] Issa Kohler-Hausmann made such an argument at this Symposium. *Policing the Police: The 2015 Legal Forum Symposium*, U. OF CHIC. LEGAL F. (Nov. 6, 2015), https://legal-forum.uchicago.edu/page/symposium [https://perma.cc/B6TQ-GEAL]; *see also* Issa Kohler-Hausmann, *Detecting Discrimination In Policing (Or, The Dangers of Counterfactual Causal Thinking . . .)*, BALKINIZATION (Aug. 13, 2015), http://balkin.blogspot.com/2015/08/detecting-discrimination-in-policing-or.html [https://perma.cc/EXE8-5BSE].

[64] Greiner & Rubin, *supra* note 62, at 775.

[65] BLANK ET AL., *supra* note 37, at 225–27.

disparity that appear to be explained by other "race-neutral" variables may ultimately be the product of the many other ways racial discrimination has divided and constructed our society.[66]

Beyond these conceptual issues, there remain *practical* challenges associated with causal inference about the "effects of race" even in the narrower sense of discrimination by a particular decision-maker. A key problem is that unlike other "treatments" that social scientists study, race does not vary for any given individual, and is thus hard to disentangle from other characteristics. Immutable traits generally defy observational researchers' best tools for causal inference, such as quasi-experiments exploiting shocks to a treatment. Instead, researchers must use methods—such as regression, reweighting, and matching—that share a core limitation: their ability to support causal inferences depends on the assumption that the only relevant confounding variables are those the researchers observe and include in the model. Because omitted variable bias is always possible, careful researchers often refer to the race gaps that remain after controlling for observed variables simply as "unexplained," rather than claiming proof of discrimination.[67]

Fortunately, neither policy analysis nor law requires definitive answers. For policy purposes, even analyses with weaker causal identification can help narrow down the possible causes of racial disparities; theoretically informed discussion can then guide interpretation of unexplained gaps. In civil litigation, the burden of proof normally requires that the factfinder believe that the best interpretation of the evidence is that discrimination *probably* had some effect. Non-statistical evidence of causation is routinely open to multiple interpretations, and while courts have often demanded more clarity out of statistical evidence, certainty or even near-certainty is too much to ask for.[68] Moreover, it should be unnecessary to rule out every conceivable confounding variable—as suggested above, the question really should be whether the police department's explanations for disparities hold up.[69]

---

[66] *See generally* Kohler-Hausmann, *supra* note 63.

[67] *See, e.g.*, Quillian, *supra* note 61, at 303.

[68] 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. McCleskey v. Kemp, 481 U.S. 279, 297 (1987). But it's not obvious that *McCleskey*'s reasoning applies to policing at all, and it does not apply 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.

[69] 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

In the remainder of this discussion, I assume that strong causal identification is the ideal goal of research on police racial discrimination, and this assumption drives the new proposal that I outline in Part III. However, in Part II, I also examine what we can learn from observational research that falls somewhat short of this goal.

## II. ASSESSING CURRENT EMPIRICAL APPROACHES

In this Part, I assess leading current empirical approaches to the assessment of police racial discrimination. First, I consider the use of regression or similar observational methods to try to isolate the effects of such discrimination by accounting for alternative explanations for policing disparities. Section A evaluates the (limited) utility of such methods to assess *individual-level* discrimination in initial police stops; Section B considers their use to assess *neighborhood-level* discrimination in stops, which is generally somewhat easier; and Section C addresses regression analyses of *post-stop* outcomes. In Section D, I look at a small but promising body of quasi-experimental research that seeks to exploit variations in decision-makers' ability to perceive race—an approach that, under certain assumptions, obviates the necessity to observe crime and other potential confounders. Finally, in Section E, I assess lab-experimental work on implicit racial bias, which provides strong causal identification of one psychological mechanism for police discrimination, but does not directly assess real-world disparities in treatment.

### A. Individual-Level Analyses of Racial Disparities in Stops

Suppose you wanted to conduct a regression analysis to assess potential explanations for racial disparities in a city's police stop rates—in particular, you want to know whether those disparities persist even after controlling for plausible race-neutral explanations. Ideally, what kind of data would you need to carry out the analysis, and what would you do with it? This is an instructive thought exercise to begin with before we turn to the less-than-ideal data sources that observational researchers actually have.

It is useful to think about racial disparity in policing as having two key dimensions: disparities *among neighborhoods* with different racial compositions, and disparities in the treatment of individuals *within neighborhoods*. Assuming policing varies across neighborhoods, any

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).

given individual's stop probability may be affected both by the neighborhood he is in and by his own characteristics and behaviors. To assess the way these components combine to explain a city's overall racial disparity patterns, one would want an individual-level dataset—ideally covering a large random sample of individuals in the city—and you would want it to contain each individual's home address (and possibly work address), so that you could also include variables for neighborhood-level characteristics.[70]

For each individual in the sample, you would first need to know his race and whether and how often the police stopped him in the time period covered. Second, you would need to think about potential race-neutral explanations for stop patterns. Because police typically explain policing disparities by reference to differences in criminal conduct, you would want rich data on the individual's behavior over some fixed time period. You would want to know what crimes the person committed over the time period, how often, and whether they were committed in public. You would want to know more generally how much time the individual spent out in public, potentially subjecting himself to a stop. The *most* important behavioral information you would want is how often the person engaged publicly in behavior that—while not necessarily criminal—could help to produce reasonable suspicion for a stop.

Ideally, you would want to break all of this behavioral data down to a fine-grained temporal level (e.g., person-days or even person-hours) so that you could estimate the probability of being stopped conditional on being engaged in crime (or suspicious activity) *at that time*. You would also want crime rates at the neighborhood level, since police typically point to neighborhood crime-rate differences as an explanation for policing disparities, and since "high crime area" is also a factor that can contribute to constitutionally reasonable suspicion.[71] Besides crime, you might also want to disentangle the roles of other individual and neighborhood differences—for example, to distinguish racial from socioeconomic discrimination, you might want socioeconomic data about neighborhoods or individuals.

---

[70] If you wanted to separate out the effect of the individual's race from the effect of neighborhood racial composition, you'd include both variables separately in the regression. Leaving the neighborhood-composition variable out would effectively combine the two components. If you wanted to focus the analysis *only* on *intra*-neighborhood disparities, you could control for what neighborhood each individual is in (neighborhood fixed effects), which effectively controls for all inter-neighborhood differences simultaneously. However, the use of neighborhood fixed effects will filter out any police racial discrimination that occurs at the inter-neighborhood level, based on racial composition.

[71] Illinois v. Wardlow, 528 U.S. 119, 124 (2000).

If you had all this information, you could estimate the probability of being stopped in a given time period conditional on actual criminal activity, other external indicia of suspiciousness, and the individual's and neighborhood's fixed characteristics. You could use regression or a similar method (like matching or reweighting) to estimate the effect of the individual's race and his neighborhood's racial composition on stop probability, holding all other factors constant. Such a regression would effectively control for the variables that police departments typically point to as explaining policing disparities (plus perhaps other potential confounders). Although omitted variable bias is always a potential issue in this kind of analysis, if your dataset were rich enough, it would be quite plausible to attribute remaining unexplained racial gaps to police racial discrimination.

The problem is that most policing-disparity researchers do not have anything remotely approaching this dream data. While some neighborhood-level information is often available (more on this below), researchers generally have no access to *any* individual-level data on a random sample of the public. Rather, they have data from the criminal justice system, which covers only individuals the police have interacted with, and is far more limited in scope. For example, some police departments require officers to fill out forms documenting every pedestrian stop. These forms typically include basic demographic information, location, and the stated reasons for the stop, as well as the result of the stop (for example, whether an arrest was made).[72] If researchers can access these forms, they can analyze the racial- and neighborhood-level distributions of police stops.

But *explaining* those distributions is much harder. First, the forms do not give us any objective information about what stopped pedestrians were doing prior to the stop. The forms tell us only what the police officers wrote down, which may be a post hoc rationalization for the stop. In a study of whether stops are discriminatory, it is certainly not sound to assume that the very police officers being studied always record pedestrian behavior in an accurate, nondiscriminatory way. Second, and even more problematically, the forms provide no information whatsoever about persons who were not stopped by the police (or, for that matter, information about what the stoppees were doing at any time in the study period other than the moment they were stopped). So taken alone, the forms do not facilitate the kind of individual-level analysis discussed above.

---

[72] *See, e.g.*, Robin S. Engel et al., *Citizens' Demeanor, Race, and Traffic Stops*, *in* RACE, ETHNICITY, AND POLICING, *supra* note 8, at 287, 289.

Nor, typically, is there any other available source of individual-level information about criminal behavior (much less *suspicious* but noncriminal behavior) that could be linked to police outcome data. In terms of public records, at most, one might imagine being able to link some sort of public database that includes citizens' residences and demographics with their official criminal records; obtaining access to even this sort of data would be difficult. In any event, official criminal records are a rather poor proxy for actual criminality. After all, the vast majority of crimes are never detected and never enter the criminal justice system. For example, surveys suggest there are hundreds of millions of drug crimes in the U.S. each year, but only about 1.5 million drug arrests.[73] Even most reported violent and property crimes go unsolved.[74] Moreover, even for cases that are in the system, we again lack objective conduct measures—we know only what people were arrested for, convicted of, and so forth. But again, these measures may themselves be affected by police racial bias, and therefore treating them as conduct measures would introduce a troubling circularity to the analysis.

For these reasons, researchers have rarely been able to conduct individual-level analyses of racial disparities in the likelihood of police interactions conditional on behavior and other potential confounders. There are, however, a couple of potentially promising ways that researchers could collect data to support such analyses. First, researchers could conduct surveys about criminal conduct and police interactions. A survey could, in principle, ask individuals for much of the individual-level information described above. Respondents could plausibly be expected to provide rough estimates of the amounts of time each day that they typically spend engaged in various activities, to recall specific police interactions reasonably accurately, and to describe their participation in crimes, including how often they carry contraband.

---

[73] *See* OFF. OF NAT'L DRUG CONTROL POL'Y, FACT SHEET: 2010 NATIONAL SURVEY ON DRUG USE AND HEALTH (Sept. 2011), https://www.whitehouse.gov/sites/default/files/ondcp/Fact_Sheets/nsduh_fact_sheet_9-7-11_0.pdf [https://perma.cc/BXE9-HP2F] (use rates). For arrest rates, see *Arrest Data Analysis Tool*, *supra* note 1. *See also Impaired Driving: Get the Facts*, CTRS. FOR DISEASE CONTROL AND PREVENTION, http://www.cdc.gov/motorvehiclesafety/impaired_driving/impaired-drv_factsheet.html [https://perma.cc/SJR7-25Y8] (last visited Oct. 2, 2016) (finding one percent arrested out of 112 million self-reported drunk-driving instances each year).

[74] Nationally, approximately twenty percent of reported property crimes and forty-five percent of reported violent crimes are cleared. FBI, CRIME IN THE UNITED STATES 2014, OFFENSES CLEARED (Fall 2015), https://www.fbi.gov/about-us/cjis/ucr/crime-in-the-u.s/2014/crime-in-the-u.s.-2014/offenses-known-to-law-enforcement/clearances/man/clearances.pdf [https://perma.cc/P88U-6KN3].

A number of large national surveys already collect detailed information about criminal incidents, including demographic information. Some surveys, such as the Census Bureau's National Crime Victimization Survey (NCVS),[75] gather data from victims, including information on police disposition of incidents. This approach allows assessment of racial disparities in police response to reported crimes, but it is not useful for victimless crimes such as drug crimes, and provides no information about how police respond to individuals who are *not* committing crimes. Other surveys ask individuals to self-report crimes, like drug use[76] (and occasionally drug delivery).[77] Self-report surveys are potentially a good match for the kinds of analysis described above, because they produce individual-level samples including people with no criminal conduct and with no police contacts.

Such surveys do not always also ask about police interactions, but they could, and some do. A few researchers have used local youth-cohort surveys that ask such questions to assess racial and other disparities.[78] The National Survey on Drug Use and Health (NSDUH)[79] asks about the respondent's arrest history in the past year, although this variable has often been ignored in studies using the data.[80] Useful

---

[75] U.S. CENSUS BUREAU (ON BEHALF OF THE BUREAU OF JUST. STATISTICS), NATIONAL CRIME VICTIMIZATION SURVEY (2014), http://www.bjs.gov/content/pub/pdf/ncvs2_2014.pdf [https://perma.cc/YF4N-Q27G].

[76] *See* LLOYD D. JOHNSTON ET AL., NAT'L INST. OF HEALTH, MONITORING THE FUTURE, NATIONAL SURVEY RESULTS ON DRUG USE, 1975–2013: VOLUME I, 9 (2014), http://monitoringthefuture.org//pubs/monographs/mtf-vol1_2013.pdf [https://perma.cc/D5D7-BSR4]; DEP'T HEALTH & HUMAN SERVS., RESULTS FROM THE 2013 NATIONAL SURVEY ON DRUG USE AND HEALTH: SUMMARY OF NATIONAL FINDINGS 26, http://www.samhsa.gov/data/sites/default/files/NSDUHresultsPDFWHTML2013/Web/NSDUHresults2013.pdf [https://perma.cc/L9N8-37UE]. Drug use surveys find only minor racial differences, which scholars often contrast with large race gaps in drug arrests. *E.g.*, MICHELLE ALEXANDER, THE NEW JIM CROW 6–7 (2010); Yonette F. Thomas, *The Social Epidemiology of Drug Abuse*, 32 AM. J. PREV. MED. S144 (2007); *see also* 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).

[77] Richard S. Frase, *What Explains Persistent Racial Disproportionality in Minnesota's Prison and Jail Populations?*, 38 CRIME & JUST. 201, 239–40 (2009) (finding somewhat higher drug sales rates among blacks); KATHERINE BECKETT ET AL., ACLU, RACE AND DRUG LAW ENFORCEMENT IN SEATTLE 42 (2008) https://www.aclu.org/files/assets/race20and20drug20law20enforcement20in20seattle_20081.pdf [https://perma.cc/NX2T-2XDG] (using data on drug delivery from study conducted at needle exchange).

[78] David S. Kirk, *The Neighborhood Context of Racial and Ethnic Disparities in Arrest*, 45 DEMOGRAPHY 55 (2008); Robert J. Sampson, *Effects of Socioeconomic Context on Official Reaction to Juvenile Delinquency*, 51 AM. SOC. REV. 876 (1986).

[79] *See supra* note 73.

[80] Criminology studies using the NSDUH have usually used its drug use figures as an out-of-sample benchmark to compare other arrest data to. *E.g.*, Holly Nguyen & Peter Reuter, *How Risky Is Marijuana Possession? Considering the Role of Age, Race, and Gender*, 58 CRIME & DELINQ. 879 (2012). But public health and medical researchers have used the NSDUH arrest data to assess racial disparities; such research often is not framed in "policing disparity" terms per se (instead

expansions could include questions about other interactions such as stops, frisks, and searches; ask more comprehensive questions about non-drug criminal conduct; and sample persons recently incarcerated (who are currently excluded, potentially introducing sample selection bias).[81]

Still, there are limitations to this approach. Surveys are expensive, and the existing large national surveys' samples are not designed to produce valid *local* estimates, so collecting information about particular police departments' practices would require substantial new undertakings.[82] In addition, crime surveys raise concerns about accuracy,[83] and these would be exacerbated if they asked for very fine-grained information—for example, detailed accounts of each day's activities. Moreover, even if individuals report actual criminal conduct accurately, it is unrealistic to expect individuals to accurately self-appraise and remember whether they were at any given time engaging in "suspicious" conduct—moving "furtively," for instance. Rather than measuring such factors directly, researchers would realistically have to rely on an assumption that respondents' reports of *actual* criminal conduct are a good proxy for behavior giving rise to reasonable *suspicion* of criminal conduct—or at least that they are an equally good proxy across racial groups.

Another possible approach is for researchers to attempt to directly observe the criminal or suspicious behavior of individuals, as well as whether police stop them. A few prior policing-disparity studies have at least sought to observe the former. The seminal example was John Lamberth's 1994 New Jersey Turnpike study, in which researchers drove just over the speed limit and observed the drivers who passed

---

treating arrest as simply a negative health/life outcome). Pacek et al. find substantial racial disparities in the probability of arrest conditional on "disordered" marijuana use. Lauren R. Pacek et al., *Race/Ethnicity Differences Between Alcohol, Marijuana, and Co-Occurring Alcohol and Marijuana Use Disorders and Their Association with Public Health and Social Problems Using a National Sample*, 21 AM. J. ADDICT. 435, 437 (2012). Burns et al. use the data to draw share comparisons between drug buys, drug use, and drug arrests. Rachel M. Burns et al., *Statistics on Cannabis Users Skew Perceptions of Cannabis Use*, 2013 FRONTIERS IN PSYCHOL. 4 tbl.2.

[81] Under current NSDUH methodology, recently-incarcerated persons are not surveyed, which may introduce sample selection bias. If black drug arrestees are more likely to be incarcerated (as many studies indicate), excluding incarcerated persons would downward-bias estimates of racial disparity in arrest rates.

[82] BRIAN WIERSEMA, NATIONAL CONSORTIUM ON VIOLENCE RESEARCH, AREA-IDENTIFIED NATIONAL CRIME VICTIMIZATION SURVEY DATA 1 (1999); 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).

[83] *See* BARRY SPUNT, SELF-REPORT SURVEYS, 4 ENCYCLOPEDIA OF CRIME AND PUNISHMENT 1465, 1465–67 (David Levinson ed. 2002); Arthur H. Garrison, *Disproportionate Minority Arrest: A Note on What Has Been Said and How It Fits Together*, 23 NEW ENG. J. ON CRIM. & CIV. CONFINEMENT 29, 42–45 (1997).

them.[84] Subsequent highway studies have used radar.[85] Other researchers have physically observed traffic violations on city streets,[86] or used video cameras in a store to observe shoplifting.[87] In none of these studies, however, did researchers seek to observe policing outcomes for the same individuals being observed. That is presumably because doing so would have massively magnified projects that were already resource-intensive—police stop only a tiny percentage of speeders, for instance, so obtaining a sample that provided sufficient statistical power to study stop outcomes would require having observers in place for a very long time.

That being said, in some contexts it might be possible to physically observe all the necessary data for a sufficient sample. For example, such studies could be carried out at immigration, airport security, or other law enforcement checkpoints (optimally with the agency's cooperation). Researchers could record the behavior of individuals passing through, demographics, whether the individual has companions, flight origin and destination, number and type of bags, plus any computer database information that agents observe when they run the traveler's identification through the system. They could then record what the agents do: for example, diversion for additional searches. This approach would be a straightforward expansion of self-studies that agencies have already carried out.[88]

Still, the prospects of this approach are limited to contexts where both the individual conduct and the police conduct in question are

---

[84] State v. Soto, 734 A.2d 350, 351 (N.J. Super. Ct. 1996); 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 1988 THROUGH 1991 (Nov. 11, 1994).

[85] *See* James E. Lange et al., *Testing the Racial Profiling Hypothesis for Seemingly Disparate Stops on the New Jersey Turnpike*, 22 JUST. Q. 193, 211–12 (2005); ROBIN S. ENGEL ET AL., PENNSYLVANIA STATE POLICE, PROJECT ON POLICE-CITIZEN CONTACTS, YEAR 2 FINAL REPORT (MAY 2003–APRIL 2004) 64–65, 110 (2005).

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

[87] Dean A. Dabney et al., *Who Actually Steals? A Study of Covertly Observed Shoplifters*, 21 JUST. Q. 693, 711 (2004). One concern about this study is that its analysis confusingly estimates shoplifting rates after controlling for behaviors (e.g., product-tampering) that seem to be part of the shoplifting conduct itself.

[88] Federal agencies have conducted studies designed to produce benchmarks for comparisons to agency data on the demographic distribution of persons passing through checkpoint stops. BUREAU OF JUSTICE STATISTICS, NCJ 196855, ASSESSING MEASUREMENT TECHNIQUES FOR IDENTIFYING RACE, ETHNICITY, AND GENDER: OBSERVATION-BASED DATA COLLECTION IN AIRPORTS AND AT IMMIGRATION CHECKPOINTS 1 (2003), http://www.bjs.gov/content/pub/pdf/amtireg.pdf [https://perma.cc/QP6A-LMT6]. The border control checkpoint study merely recorded the race of those passing through (a population benchmark), while the airport security checkpoint study recorded some additional information such as gender, age, and number of carry-ons.

expected to take place in known places and times where researchers or cameras can be stationed. Most crime and most police interactions, however, are less predictable, and surveys will be the only plausible source of individual-level data. In practice, given the demands of either method, we can expect that in most policing contexts, it will remain very difficult to carry out observational studies of policing disparities that seek to measure and control for differences in individual behavior.

### B. Neighborhood-Level Analyses

Suppose you had a more limited objective: to assess only the *inter-neighborhood* dimension of possible police racial discrimination. Do police treat neighborhoods of color differently because of their racial composition? Regression analysis could seek to answer this question by regressing neighborhood stop rates on measures of racial composition (for instance, the black population share) as well as plausible neighborhood-level confounding variables—that is, nonracial reasons that police might treat different neighborhoods differently.

It *is* typically possible to construct datasets to support this kind of analysis, and researchers have done so. An important example was the study of precinct-level disparities presented by the plaintiffs' expert, Jeffrey Fagan, in the *Floyd* stop-and-frisk case against the NYPD.[89] Fagan regressed precinct stop rates on racial groups' population shares; controls included crime complaint rates, neighborhood socioeconomic and other demographic characteristics, and size of the precinct's police force.[90] The report found that black and Hispanic population share strongly predicted stop rates,[91] and the district court agreed.[92]

Highly localized demographic and socioeconomic data are readily available (from the Census Bureau and other sources), but what about crime data? Various crime statistics can also readily be obtained at local levels (e.g., the precinct), but how well do these approximate rates of *actual* crime, or for that matter, rates of "suspicious behavior?"

---

[89] Report of Jeffrey Fagan, Ph.D., 30-34, Floyd v. City of New York, 959 F. Supp. 2d 540 (S.D.N.Y. 2013) (No. 08 Civ. 01034).

[90] *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.

[91] *Id.* at 32–34. 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 crime or other confounders, which weakens their causal identification. However, because NYPD was overwhelmingly stopping innocent people, the absence of crime controls may not be seriously problematic; *see supra* note 55.

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

Some scholars have used arrest rates to stand in for crime rates, but this really should be avoided, because it again introduces circularity.[93] Arrests are discretionary police decisions, and thus could be infected by racial discrimination. When arrests are used as a crime measure in stop-disparity studies, what is really being asked is not "Does crime explain stop disparities?" but "How do stop disparities compare to arrest disparities?"

A much better option is to control for the neighborhood's *reported* crime rates, which principally come from calls made by citizens.[94] But this approach still faces the problem that most crime is unreported—about half of violent crimes and sixty percent of property crimes, according to victim surveys,[95] while minor or victimless crimes are almost *never* reported.[96] And *reported* crime might not be a race-neutral (or neighborhood-neutral) proxy for *actual* crime, potentially biasing analyses. This concern is amplified because (due to the extreme underreporting of other crime types) reported-crime benchmarks are usually based on the FBI's serious "index crimes,"[97] violent crimes, or just homicide.[98] But racial differences in crime rates are believed to be far greater for violent crime, especially homicide, than for other crimes.[99] If police stops are driven substantially by more minor crimes,

---

[93] *E.g.*, Jeffrey A. Fagan et al., *Street Stops and Broken Windows Revisited*, *in* RACE, ETHNICITY, AND POLICING, *supra* note 8, at 318–19; 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 (2007).

[94] The FBI's Uniform Crime Reports (UCR) provide summary data for certain crimes, but only collects race information for homicides. FBI, CRIME IN THE UNITED STATES 2014, OFFENSES KNOWN TO LAW ENFORCEMENT (Fall 2015), https://www.fbi.gov/about-us/cjis/ucr/crime-in-the-u.s/2014/crime-in-the-u.s.-2014/offenses-known-to-law-enforcement/main/offenses-known-to-law-enforcement.pdf [https://perma.cc/4BMY-6BBN]. 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, VOLUME 1: DATA COLLECTION GUIDELINES 63–64 (2000), http://www2.fbi.gov/ucr/nibrs/manuals/v1all.pdf [https://perma.cc/E7K5-CYH5]. Some studies use crime-report data from local sources. *E.g.*, HOWARD P. GREENWALD ET AL., FINAL REPORT: POLICE VEHICLE STOPS IN SACRAMENTO, CALIFORNIA 16–18 (2001); GREG RIDGEWAY, RAND CORP., ANALYSIS OF RACIAL DISPARITIES IN THE NEW YORK POLICE DEPARTMENT'S STOP, QUESTION, AND FRISK PRACTICES 13 (2006).

[95] JENNIFER L. TRUMAN, BUREAU OF JUSTICE STATISTICS, NCJ 235508, CRIMINAL VICTIMIZATION, 2010, 1 (Sept. 2011), http://www.bjs.gov/content/pub/pdf/cv10.pdf [https://perma.cc/7CEQ-JVMM].

[96] Sampson & Lauritsen, *supra* note 53, at 317.

[97] The Uniform Crime Reports crime index includes murder, rape, arson, larceny, robbery, burglary, car theft, and aggravated assault. FBI, CRIME IN THE UNITED STATES 2014, EXPANDED OFFENSE DATA (Fall 2015), https://www.fbi.gov/about-us/cjis/ucr/crime-in-the-u.s/2014/crime-in-the-u.s.-2014/offenses-known-to-law-enforcement/expanded-offense/expanded-offense/expanded-offense-data.pdf [https://perma.cc/NFU8-9FXK].

[98] *E.g.*, Fagan et al., *supra* note 93, at 318–19 (using homicide); RIDGEWAY, *supra* note 94, at xii, 16–19 (using violent crime).

[99] Frase, *supra* note 77, at 238.

using reported violent-crime rates as a proxy for *all* crime risks substantially overstating the extent to which crime differences explain stop disparities across neighborhoods.

Arguably, for the purpose of assessing police discrimination, reported crime measures could be the right benchmark despite the enormous amount of crime that gets left out of them, because "the criminal justice process does not begin until the police become aware of a crime."[100] That is, if you are trying to model the inputs into police decision-making, you should not worry about the crime that the police never hear about. This point is probably overstated, because in many cities, a large fraction of policing is not report-driven at all, but proactive—for example, patrol officers walking or driving the streets looking for suspicious activity.[101] Still, surely police departments' approaches to different neighborhoods are at least *somewhat* influenced by reported crime levels. So reported-crime measures are probably essential to include in neighborhood-level analyses of police discrimination, and they are perhaps defensible proxies for unmeasured (but observable-to-police) suspicious or criminal activity as well. Researchers should, however, be cognizant of their limitations and possible biases.

### C. Studies of Post-Stop Outcomes

While individual-level analyses of disparities in police *stops* have been almost nonexistent, individual-level data have often been used to study disparities in searches, arrests, or other sanctions *among stopped persons*.[102] These studies are made possible by the data that many police departments require officers to collect on those they stop. On the surface, these studies are more straightforward than attempts to estimate stop disparities, because for each stopped individual, it appears possible to observe all the key variables: what they were stopped for, demographics, where the stop occurred, how they acted, and the outcome of the stop. However, these studies raise substantial causal inference challenges.

First, if there is racial discrimination in initial stops, the samples of stopped persons of different races differ in unobservable ways, introducing sample selection bias. For example, consider Smith and

---

[100] Garland et al., *supra* note 50, at 19–20,

[101] *See* RIDGEWAY, *supra* note 94, at 18 (finding thirty percent of NYPD stops were initiated by citizen calls or suspect descriptions).

[102] *E.g.*, Pickerill et al., *supra* note 34, at 9–19; Ridgeway & MacDonald, *supra* note 45, at 192–98; Greg Ridgeway, *Assessing the Effect of Race Bias in Post-Traffic Stop Outcomes Using Propensity Scores*, 22 J. QUANT. CRIMINOLOGY 1, 1 (2006).

Petrocelli's study of Richmond traffic stops, which found that after controlling for observable differences, stopped minority drivers were substantially *less* likely to be ticketed or arrested.[103] The authors offer two possible interpretations of their finding: first, that police discriminated in favor of minorities,[104] or second, that they sanctioned minorities less often because more minorities had been unjustifiably stopped.[105] Notice that these interpretations are effectively opposite, and choosing between them requires assumptions about stop decisions.[106]

Second, these studies have often provided good illustrations of how challenging it is to specify a model properly, including appropriate control variables but not inappropriate ones. Studies of post-stop disparities often include control variables that risk filtering out part of the police discrimination that the study is trying to measure. For example, analyses of search rates often control for whether the individual is arrested or sanctioned.[107] The apparent rationale is that arrests and sanctions reflect conduct, or that searches incident to arrest are not discretionary.[108] But arrest and sanction decisions are themselves discretionary. Moreover, some arrests result from searches, not vice-versa, or may be motivated by the desire to carry out a search. When studying search decisions, it is inappropriate to control for something that is itself shaped by the search decision; doing so likely biases estimates of unexplained disparity downward. These examples illustrate that concerns about omitted variable bias cannot simply be solved by a kitchen-sink approach to constructing a model—that is, more control variables are not always better. If one is trying to study police racial discrimination, variables likely to be influenced by that very discrimination are "bad controls."

---

[103] Michael R. Smith & Matthew Petrocelli, *Racial Profiling? A Multivariate Analysis of Traffic Stop Data*, 4 POLICE Q. 4, 18–20 (2001).

[104] *Id.* The researchers posit that officers, aware that the research study was taking place, may have altered their behavior when interacting with black drivers to avoid accusations of racism.

[105] *Id.*

[106] A similar problem arises in studies of disparities in later process stages, such as sentencing or plea-bargaining. Most such studies compound the problem by using samples consisting only of *sentenced* cases and controlling for conviction severity, failing to account for disparities in charging and plea-bargaining, in addition to police decision-making. *See* Sonja B. Starr & M. Marit Rehavi, *Mandatory Sentencing and Racial Disparity*, 123 YALE L.J. 2, 39–77 (2013) (explaining this problem).

[107] *See, e.g.*, Alpert et al., *supra* note 86, at 46–47 (controlling for whether the individual is arrested); Pickerill et al., *supra* note 34, at 9–19 (controlling for the number of violations).

[108] *See* Pickerill et al., *supra* note 34, at 15.

That said, these research design choices are not always easy. Consider the dilemma of whether to control for behavior recorded by officers on stop forms. As discussed in Part I, officers' decisions about what to write down could likewise be affected by race. Scholars have nonetheless often included officers' descriptions as control variables, implicitly treating them as accurate, exogenous descriptions of behavior rather than as discretionary decisions. For example, a Cleveland study found that arrest disparities disappeared after controlling for whether police described drivers as noncompliant or disrespectful.[109] The RAND study of NYPD's frisk, search, and sanction rates controlled for "evasiveness, . . . 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."[110]

Other studies exclude such subjective factors from their models. I believe this is the better choice, at least for factors that are easy to manipulate. However, it does risk omitted-variable bias, because at least sometimes the officers' descriptions presumably *will* be grounded in actual conduct. There are no perfect choices. Ideally, researchers should investigate whether their estimates are affected by alternative choices on difficult model specification questions. Careful observational studies of stops as well as post-stop outcomes are potentially informative, but researchers must remember their limits.

### D. Exploiting Variations in Enforcers' Information About Race

Alternatively, instead of simply trying to measure and control for all confounding variables, researchers sometimes look for quasi-experimental approaches—that is, approaches that exploit natural shocks to the treatment variable. While race itself is not subject to such shocks, officers' *perception* of race sometimes is, and a few studies have taken advantage of this fact. If racial disparity increases when officers are likelier to know an individual's race, this increase can reasonably be attributed to discrimination, assuming this change in knowledge is not also accompanied by changes in other relevant factors.

Several studies have compared officers' traffic enforcement decisions to truly race-blind mechanisms: traffic-camera citations[111]

---

[109] Engel et al., *supra* note 72, at 297–99.

[110] Ridgeway, *supra* note 102, at 34–35.

[111] MONTGOMERY COUNTY DEP'T OF POLICE, TRAFFIC STOP DATA ANALYSIS: THIRD REPORT (2002).

and citations issued via aerial surveillance.[112] These are very informative designs, analogous to strong studies on other discrimination questions—for example, research demonstrating increased hiring of women when orchestras adopted blind auditions.[113] A limitation is that race information is not the only difference between the human and the automated decision processes: automated and aerial enforcement target only a subset of the violation types that human officers respond to.[114]

Several traffic-stop studies have exploited variation in race information known to officers, on the theory that a driver's race is harder to see at night. The studies compare stops at the same clock time but on either side of Daylight Savings Time transitions, which changes whether night has fallen. The intuition is that if racial disparity is driven by discrimination, it should be reduced at night. Studies in Portland and Cincinnati found no such reduction, concluding that disparities were not caused by racial discrimination.[115] Studies in Minneapolis and Syracuse reached the opposite conclusion; the Syracuse study, unlike the others, accounted for variations in artificial light.[116]

One interpretive problem is that police and drivers may behave differently when it is dark for reasons that have nothing to do with the reduction in police ability to perceive drivers' race. If so, the research design does not allow these race-neutral reasons to be disentangled from the reduced race information. Darkness certainly affects driving behavior, and could also affect police tactics, or police perceptions of black criminality.[117]

Still, these studies are very clever and represent some of the strongest observational research on policing disparity, and the general strategy of exploiting variations in race information is promising.

---

[112] 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*, ANN. MTG. ACAD. CRIM. JUST. SCI. (2001).

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

[114] *See* Ridgeway & MacDonald, *supra* note 45, at 183 (citing these studies and raising this concern).

[115] 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., RAND CORP., POLICE–COMMUNITY RELATIONS IN CINCINNATI: YEAR THREE EVALUATION REPORT 27 (2007).

[116] Joseph A. Ritter & David Bael, *Detecting Racial Profiling in Minneapolis Traffic Stops: A New Approach*, CURA REPORTER (Spring/Summer 2009) at 11–17; William C. Horrace & Shawn M. Rohlin, *How Dark Is Dark? Bright Lights, Big City, Racial Profiling*, 98 REV. ECON. & STATS. 226, 231 (May 2016).

[117] 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 J. 24, 35 (2009).

However, the potential of this research design is limited to narrow contexts: those in which enforcement decisions can be made without close-range observation of suspects.

### E. Lab Experiments on Implicit Biases

Aside from these observational approaches, many lab experiments demonstrate the prevalence of "implicit racial bias," including association of blackness with criminality.[118] For example, Eberhardt et al. showed that police subjects who were primed subconsciously with crime-related images paid disproportionate attention to black faces.[119] A subset of this literature tests "shooter bias," using computer simulations; subjects are asked to "shoot" armed characters. These tests find that players pick the right response faster if the image matches stereotypes (e.g., armed black characters).[120]

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 test the association between blackness and weapons, which is prevalent: one analysis found that seventy-two percent of respondents showed this association, and only nine percent showed the reverse.[121] 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 be *less* biased than average (although this is speculative, of course), and most respondents are presumably trying to achieve an "unbiased" score.

This research strongly indicates that implicit racial bias is prevalent among police—but not limited to them. Police and civilian subjects score similarly, and on some tasks police make fewer mistakes

---

[118] *See, e.g.*, B. Keith Payne, *Weapon Bias*, 15 CURRENT DIR. IN PSYCH. SCI. 287 (2006) (reviewing literature); Quillian, *supra* note 61, at 314–20 (same); CHERYL STAATS, KIRWAN INSTITUTE, STATE OF THE SCIENCE: IMPLICIT BIAS REVIEW 2014, http://kirwaninstitute.osu.edu/wp-content/uploads/2014/03/2014-implicit-bias.pdf [https://perma.cc/J5AP-BFY4] (same).

[119] Jennifer L. Eberhardt et al., *Seeing Black: Race, Crime, and Visual Processing*, 87 J. PERSONALITY & SOC. PSYCH. 876, 885–88 (2004) (also finding that crime-primed officers were more likely to wrongly pick a more racially "stereotypical" black face out of a lineup); Heather M Kleider et al., *Looking Like a Criminal*, 40 MEMORY & COGNITION 1200, 1200 (2012) (reaching similar findings with student subjects).

[120] 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 (2003).

[121] Brian A. Nosek et al., *Pervasiveness and Correlates of Implicit Attitudes and Stereotypes*, 2007 EUR. REV. SOC. PSYCH. 1, 20 (2007).

overall.[122] 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."[123] Surveys have also shown widespread tendencies to *explicitly* associate blackness with criminality,[124] and overt endorsement of racial discrimination among a small but nontrivial subset of white respondents.[125]

The great unknown is how implicit bias affects real-world police decisions.[126] A key next step is to link implicit bias scores to real-world policing outcomes—for example, within police departments, do individual officers' scores tend to predict racial disparity in their stop rates? If so, it would support police efforts to reduce implicit bias, perhaps via "debiasing" trainings or by using implicit bias testing in hiring.[127] If not, it might suggest that the recent focus by many departments and researchers on implicit bias (rather than explicit bias or behavior) is misguided. Such studies would have limitations: while the tests are controlled experiments, using their results to explain real-world outcomes involves the usual causal-inference challenges of observational research. For example, an officer's experiences could influence both her implicit bias scores and her stop practices. Still, such studies could be a promising approach to assessing one plausible mechanism for disparities.

## III. AUDITING: A FIELD EXPERIMENTAL APPROACH

As the review in the last Part suggests, evaluating police racial discrimination is truly difficult, and despite decades of serious effort, our existing tools have serious shortcomings. Most observational methods suffer from data shortcomings and causal inference challenges; quasi-experimental methods are of limited applicability;

---

[122] Joshua Correll et al., *Across the Thin Blue Line: Police Officers and Racial Bias in the Decision to Shoot*, 92 J. PERSONALITY & SOC. PSYCH. 1006; *see generally* Correll et al., *supra* note 120, at 206.

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

[124] James D. Unnever, *Race, Crime, and Public Opinion*, *in* THE OXFORD HANDBOOK OF ETHNICITY, CRIME, AND IMMIGRATION 70, 71 (Sandra M. Bucerius & Michael Tonry eds. 2014).

[125] *See, e.g.*, Frank Newport, *In U.S., 87% Approve of Black-White Marriage, vs. 4% in 1958*, GALLUP (July 25, 2013), http://www.gallup.com/poll/163697/approve-marriage-blacks-whites.aspx [https://perma.cc/Y825-JAL4] (reporting 2013 poll showing that only eighty-four percent of white Americans approve of interracial marriage).

[126] *E.g.*, BLANK ET AL., *supra* note 37, at 72 ("[L]aboratory effects . . . can rarely tell us the extent to which naturally observed disparities are the result of discrimination.").

[127] Caution is appropriate; due to random variation, any one individual's score may not mean much. However, this problem could probably be mitigated by using more extensive or repeated testing.

and lab research has uncertain implications for the "real world." Accordingly, I propose a new method to supplement the existing toolkit: the use of "testers," also known as "auditing." While the term "auditing" has various meanings in other contexts, in antidiscrimination research it is usually used to refer to field studies that compare the treatment of paired individuals ("testers") who are similar but for a specific characteristic such as race. Such methods are used often in discrimination research and civil rights law enforcement in areas such as employment, housing, and lending. I propose using 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.

### A. 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.[128] The authors matched white male testers with black male, black female, and white female counterparts based on age, education, and assessed attractiveness.[129] 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.[130] Other studies have used auditing to study housing and employment markets,[131] in addition to various other phenomena—for example, a recent study found that drivers are less willing to yield to black jaywalkers.[132]

Some studies manipulate only written information, such as employment applications,[133] student emails to professors,[134] and writing

---

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

[129] *Id.* at 306.

[130] *Id.* at 319. The evidence of gender discrimination was less clear.

[131] *E.g.*, John Yinger, *Measuring Discrimination with Fair Housing Audits: Caught in the Act*, 76 AM. ECON. REV. 881 (1986); *see* BLANK ET AL., *supra* note 37, 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 AM. ACAD. POL. & SOC. SCI. 104, 114 (2007) (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).

[132] Tara Goddard et al., *Racial Bias in Driver Yielding Behavior at Crosswalks*, 33 TRANSPORTATION RESEARCH PART F 1, 5 (2015); *see* BLANK ET AL., *supra* note 37, at 104–08 (reviewing auditing literature); Riach & Rich, *supra* note 131.

[133] *E.g.*, Marianne Bertrand & Sendhil Mullainathan, *Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination*, 94 AM. ECON. REV.

samples.[135] Such designs allow true experimental manipulation of race and gender, which in-person auditing does not quite achieve: one can randomize cases between testers, but one cannot 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, a criminal defendant introduced evidence from testers that he had hired to assess whether race affected Border Patrol stops. But the experiment was tiny, and the unpersuaded court observed that many conditions had not been held constant.[136] A Mexico City study used testers who committed illegal left turns to test socioeconomic status effects on police demands for bribes.[137] Another study focused on private party suspicions of crime, testing store clerks' reactions to white and black shoppers.[138] 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.[139]

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 held that testers could have standing.[140] 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.[141] Testers have also brought challenges to lending

---

991 (2004); Pager, *supra* note 131, at 942–43 (reviewing studies).

[134] Katherine L. Milkman et al., *What Happens Before? A Field Experiment Exploring How Pay and Representation Differentially Shape Bias on the Pathway Into Organizations*, 100 J. APPLIED PSYCHOL. 1678, 1696–98 (2015).

[135] Arin N. Reeves, Nextions, Written in Black & White: Exploring Confirmation Bias in Racialized Perceptions of Writing Skills 4–6 (2014), http://www.nextions.com/wp-content/files_mf/1446822647201404011 4WritteninBlackandWhiteYPS.pdf [https://perma.cc/9DHV-N2AN].

[136] United States v. Beasley, 36 F.3d 1106, No. 94-2026, No. 94-2065 (10th Cir. 1994) (unpublished table decision).

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

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

[139] *What Would You Do? (Bike Thief)* (ABC television broadcast 2010), http://www.youtube.com/watch?v=S0kV_b3IK9M [https://perma.cc/Z7ZS-E8N7].

[140] Evers v. Dwyer, 358 U.S. 202 (1958) (per curiam).

[141] *See* Havens Realty Corp. v. Coleman, 455 U.S. 363 (1982) (holding that a tester could have standing if he could show actual injury); Michael Selmi, *Public vs. Private Enforcement of Civil Rights*, 45 U.C.L.A. L. REV. 1401, 1426 (1998); MARGERY A. TURNER ET AL., THE URBAN INSTITUTE, DISCRIMINATION IN METROPOLITAN HOUSING MARKETS: NATIONAL RESULTS FROM PHASE I HDS 2000 (2002).

discrimination,[142] and the Equal Employment Opportunity Commission has endorsed their use to challenge hiring discrimination, though few cases have been brought.[143]

## B. Auditing the Police: Key Research Design Considerations

Is auditing the police realistic? This has not been done before,[144] and there are some good reasons for that—but I believe these concerns can be addressed with careful research design. Here, I address several objectives that researchers must balance: safety, legality, importance, methodological rigor, statistical power, and cost concerns.

### 1. Safety.

A paramount concern that will limit the potential scope of this approach 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.

---

[142] Steve Tomkowiak, *Using Testing Evidence in Mortgage Lending Discrimination Cases*, 41 URB. LAW. 319, 326–36 (2009).

[143] EQUAL EMP'T OPPORTUNITY COMM'N., DEC. NO. 915.002, ENFORCEMENT GUIDANCE: WHETHER "TESTERS" CAN FILE CHARGES AND LITIGATE CLAIMS OF EMPLOYMENT DISCRIMINATION (May 22, 1996); Marc Bendick, Jr. & Ana P. Nunes, *Developing the Research Basis for Controlling Bias in Hiring*, 68 J. SOC. ISSUES 238, 256 (2012).

[144] Indeed, aside from the *Beasley* defendant's effort, *see supra* note 136, it has hardly been suggested. One scholarly piece and one news article each give the idea a sentence or two. Pamela S. Karlan, *Race, 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 (Apr. 30, 2014), https://www.washingtonpost.com/news/wonk/wp/2014/04/30/it-is-exceptionally-hard-to-get-good-data-on-racial-bias-in-policing/ [https://perma.cc/EE24-NTP3]. Reviews of methods for studying racial profiling omit it; for example, Blank et al. don't mention auditing in their chapter on police, even though they endorse it for other contexts like housing. BLANK ET AL., *supra* note 37, at 103–17, 186–202.

2. Legality.

The criminal law constrains staging of actual crimes, lying to the police, and recording of interactions.[145] This is another advantage of governmental involvement. Undercover police routinely participate in otherwise-criminal activity and enjoy effective immunity from prosecution.[146] Private testers cannot 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,[147] the jaywalking study,[148] and the Mexico City bribery study.[149] Most of the designs proposed below involve no lawbreaking or lying, just potentially suspicious activity.

3. 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,[150] and may be a method of expanding the surveillance "net," exposing arrestees to more police interactions in the future.[151]

4. Methodological rigor.

The most obvious requirement for effective auditing is that the deception must work. The interaction should thus be quite ordinary, brief, and forgettable. Observations should be distributed across police

---

[145] 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 (2012), http://www.rcfp.org/rcfp/orders/docs/RECORDING.pdf [https://perma.cc/3FUZ-UX8X]. 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 have found only a right to *openly* record the police, Crawford v. Geiger, 996 F. Supp. 2d 603, 614–15 (N.D. Ohio 2014).

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

[147] LAMBERTH, *supra* note 84.

[148] Goddard et al., *supra* note 132.

[149] Fried et al., *supra* note 137.

[150] *See supra* note 16 (discussing Whren v. United States, 517 U.S. 806 (1996)). Arrestees may be searched without warrants.

[151] Issa Kohler-Hausmann, *Managerial Justice and Mass Misdemeanors*, 66 STAN. L. REV. 611, 632–33, 639 (2014).

beats and shifts and across time, so individual officers are unlikely to notice patterns.

The primary threat to causal inference from auditing studies is tester heterogeneity,[152] so testers should be matched carefully.[153] 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 is 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),[154] 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.

5. Statistical power and cost.

The sample size must provide sufficient statistical power to produce reasonably precise estimates.[155] Ideally, this means at least hundreds of observations[156]—a plausible number (large cities have thousands of officers), provided the tests are spread across beats and shifts.[157] Many published auditing studies have much smaller samples,

---

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

[153] *See* Pager, *supra* note 131, 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 131, at F483–F484.

[154] *See, e.g.*, Ayres & Siegelman, *supra* note 128, at 307 (using blind testers).

[155] 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.

[156] Sample-size calculators are widely available; they require assumptions about effect size. For example, if one seeks eighty percent power with a ninety-five confidence level, assuming the true probabilities for the two groups are thirty percent and forty percent respectively, a common power formula requires a total sample size of 708. *See Power (Sample Size) Calculators*, SEALED ENVELOPE, https://www.sealedenvelope.com/power/binary-superiority/ [https://perma.cc/GP73-55UV] (last visited Oct. 2, 2016). If the probabilities were thirty percent and fifty percent, the sample size required would be smaller (182).

[157] For example, the Chicago police department has 279 distinct beats, each patrolled by eight or nine officers. *Beat Officers*, CHI. POLICE DEP'T, http://home.chicagopolice.org/get-involved-with-caps/how-caps-works/beat-officers/ [https://perma.cc/AG4B-YYNR].

allowing them to detect only large effects, and even then, imprecisely.[158] 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.

### C. Possible Research Designs

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

*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 are asked what is 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,[159] 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.[160] Outcomes to be measured could include time elapsed during an encounter or the rate of diversion for

---

[158] *E.g.*, Fried et al., *supra* note 137, at 83 (42 tests); Schreer et al., *supra* note 138, at 1438 (thirty-three tests, six stores).

[159] *Supra* note 139.

[160] *Cf. supra* note 88 and accompanying text (discussing observational checkpoint studies). Testers would have some advantages versus other approaches to studying checkpoints, in that one can easily hold constant subtle behaviors or differences in verbal responses that might be difficult for observational researchers to measure and code.

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.[161] The test should focus on initial intake, with a mechanism for stopping the ensuing investigations.

## D. 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 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.[162]

The downside is that 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

---

[161] *See, e.g.*, RICHARD J. DAVIS, 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 their phone-based auditing study).

[162] *See* Quillian, *supra* note 61, at 303 ("[A]udit studies often are the best method for measuring . . . discrimination.").

designs above would test disparities *within neighborhoods* (or at checkpoints), and would miss differences driven by neighborhood racial composition. Auditing is best designed to address *intra*-neighborhood disparities, which is an important limitation, although it bears emphasis that even if this were *all* it was good for, it would fill an important gap. As Part II illustrated, regression methods are a reasonably effective tool for estimating inter-neighborhood disparities; it is the disparate treatment of individuals within neighborhoods that is the most difficult to get at using observational methods, because of the absence of individual-level data.

The auditing design could, in any event, be extended to test the effects of the neighborhood or other race-correlated variables and their interaction with individual race—for example, by changing the same testers' clothing or sending them to different neighborhoods. An advantage over observational studies of inter-neighborhood disparities is that this approach could rule out, as an explanation for those disparities, inter-neighborhood differences in the behavior of the individuals being approached. That is, if testers act the same in every neighborhood, systematic differences in their treatment across neighborhoods would provide strong evidence that it is not the testers, but an actual difference in police approaches. Still, even if auditing reveals inter-neighborhood disparities, it would not necessarily mean that that difference is *because of* neighborhood racial composition.[163] 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.[164] 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.

### E. 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,[165]

---

[163] Researchers could try to account for other neighborhood differences by adding the same sorts of controls (e.g., reported crime rate differences) that observational studies do.

[164] *Cf.* Heckman, *supra* note 152, at 102–11 (criticizing employment audit studies for not estimating market discrimination).

[165] *See* Selmi, *supra* note 141, at 1427; Alex Young K. Oh, *Using Employment Testers to Detect Discrimination: An Ethical and Legal Analysis*, 7 GEO. J. LEGAL ETHICS 473, 480 (1993) (citing

and certainly may be in this context as well, but there are countervailing political pressures. In surveys, large majorities oppose racial profiling.[166] The Civil Rights Division of the U.S. Department of Justice has a strong interest in the issue and in police abuses generally,[167] and the issue has been a high overall priority following recent events.[168] Outside-agency auditing would lose some of the advantages of police-department self-monitoring (for example, access to internal data), but outside auditors could still employ 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, the New Jersey Attorney General's office carried out a major benchmarking study under a settlement with the U.S. Department of Justice.[169] Outside monitors have been appointed for numerous police departments, often under consent decrees.[170]

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. Still, while many departments would doubtless reject the idea, the 18,000 law enforcement agencies in the U.S. are not monolithic. Typically, agency heads are political appointees, and there is no reason to assume that all cities' political

---

employer fears of tester litigation).

[166] Emily Eakins, *Poll: 70% of Americans Oppose Racial Profiling by the Police*, REASON FOUNDATION (Oct. 14, 2014), http://reason.com/poll/2014/10/14/poll-70-of-americans-oppose-racial-profi [https://perma.cc/84GM-CQ2E].

[167] *See 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 [https://perma.cc/777J-DAQM]; U.S. DEP'T OF JUSTICE, CIVIL RIGHTS DIV., GUIDANCE FOR FEDERAL LAW ENFORCEMENT AGENCIES REGARDING THE USE OF RACE, ETHNICITY, GENDER, NATIONAL ORIGIN, RELIGION, SEXUAL ORIENTATION, OR GENDER IDENTITY (2014), https://www.justice.gov/sites/default/files/ag/pages/attachments/2014/12/08/use-of-race-policy.pdf [https://perma.cc/W5LX-JBBE].

[168] Press Release, U.S. Dep't of Justice, Statement by Attorney General Eric Holder on Latest Developments in Ferguson, Missouri (Aug. 14, 2014), https://www.justice.gov/opa/pr/statement-attorney-general-eric-holder-latest-developments-ferguson-missouri [https://perma.cc/5PYG-XES9].

[169] *See* Lange et al., *supra* note 85, at 196–97.

[170] *See, e.g.*, *Agency Profiles*, NAT'L ASSOC. FOR CIVILIAN OVERSIGHT OF LAW ENF'T, http://www.nacole.org/agency_profiles [https://perma.cc/SP6M-3XT6] (providing detailed profiles of civilian oversight agencies); *see also* Sanchez et al., *supra* note 32 (describing the three-year monitoring plan adopted by New York City as part of the settlement reached in the *Floyd* litigation); Barbara Attard, *Oversight of Law Enforcement Is Beneficial and Needed—Both Inside and Out*, 30 PACE L. REV. 1548, 1550 (2010).

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.[171] Some police departments have "early warning" programs to identify individual problem officers.[172] Any of these programs risk litigation or officer backlash—indeed, programs that risk getting individual officers in trouble may raise a worse risk of backlash than auditing does.[173] These risks have not precluded these programs' adoption.

There is substantial precedent for using 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.[174] Undercover agents are also often employed in police corruption investigations.[175] Several police departments (including New York and Los Angeles) regularly conduct "random integrity tests"—exposing officers to random stings.[176] Corruption is likely as embarrassing to police departments as racial discrimination is—yet these departments have carried out the corruption equivalent of auditing.

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

---

[171] *See, e.g.*, ENGEL et al., *supra* note 85; *Law Enforcement*, CTR. FOR POLICING EQUITY, http://policingequity.org/law-enforcement/ [https://perma.cc/2RHV-32X8] (describing CPE's work with police departments).

[172] Robin S. Engel & Jennifer Calnon, *Comparing Benchmark Methodologies for Police Citizen Contacts: Traffic Stop Data Collection for the Pennsylvania State Police*, 7 POLICE Q. 97, 109 (2004); *see* RIDGEWAY, *supra* note 94, at 21–30.

[173] Unions generally strongly oppose policies with potential adverse consequences for individual officers. Engel & Calnon, *supra* note 172, at 109; Kevin M. Keenan & Samuel Walker, *An Impediment to Police Officer Accountability?*, 14 B.U. PUB. INT. L.J. 185, 198–99 (2005).

[174] The term comes from military wargaming exercises. Michael K. Meehan, *Red Teaming for Law Enforcement*, 74 POLICE CHIEF 22; *see* FED. BUREAU OF INVESTIGATION, Subject Bibliography: Red Teaming, https://www.hsdl.org/?view&did=702932 [https://perma.cc/9VB6-MKUX] (collecting sources); William H. Adcox, *The Red Team: An Innovative Quality Control Practice in Facility Security*, 74 POLICE CHIEF 54 (2007) (describing "breach exercise[s]" carried out by undercover teams at protected facilities).

[175] *E.g.*, Steve Rothlein, Legal & Liability Risk Management Institute, *Conducting Integrity Tests on Law Enforcement Officers*, LEGAL LIABILITY AND RISK MANAGEMENT INSTITUTE (2010), http://www.patc.com/weeklyarticles/print/le_integrity_tests.pdf [https://perma.cc/9B3D-SH9C]; *see* Tim Prenzler & Carol Ronken, *Police Integrity Testing in Australia*, 1 CRIMINOLOGY & CRIM. JUST. 319, 319 (2001) (describing undercover integrity testing in Australia as an "essential" anticorruption tool).

[176] Rothlein, *supra* note 175; Prenzler & Ronken, *supra* note 175, at 321–23; Sanja Kutnjak Ivkovic, 93 J. CRIM. L. & CRIMINOLOGY 593, 617–19 (2003).

criticized for papering over police misconduct and corruption.[177] 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[178]—provided those outside actors are truly independent.[179] 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.[180]

If police departments are reluctant to expose themselves to liability or criticism, they could conduct internal auditing programs without publicizing results, or ask academic collaborators to publish anonymized results. To encourage self-studies, legislatures could consider creating 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.[181] The privileges apply only if, upon discovering evidence of discrimination, the lender undertakes "appropriate corrective action."[182] 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).[183]

If government involvement proves impracticable, academic researchers might be able to carry out some of the designs on their own,

---

[177] *E.g.*, Ivkovic, *supra* note 176, at 596–97.

[178] This is the modus operandi of the Center for Policing Equity, which connects researchers with police departments. *See Law Enforcement*, *supra* note 171; *see* Merrick Bobb, *Civilian Oversight of the Police in the United States*, 22 ST. LOUIS U. PUB. 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).

[179] Civilian oversight boards have often been criticized for being overly deferential to police. *E.g.*, Stephen Clarke, *Arrested Oversight: A Comparative Analysis and Case 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.

[180] *Cf.* Riach & Rich, *supra* note 131, at F483 (worrying that "consciously or unconsciously, minority applicants may be motivated to prove the existence of discrimination."); *see also* Heckman, *supra* note 152, at 104. When police are investigating police, one might worry more about the opposite concern.

[181] *See* Tomkowiak, *supra* note 142, at 325–27.

[182] *Id.*

[183] *See* Josh Jones, Note, *Behind the Shield? Law Enforcement Agencies and the Self-Critical Analysis Privilege*, 60 WASH. & LEE L. REV. 1609, 1611–14 (2003). Federal privilege legislation could be grounded in Congress's Fourteenth Amendment enforcement powers, and could perhaps extend to state courts.

although this would entail greater challenges. Academic research is governed by Institutional Review Board (IRB) oversight,[184] 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).[185] Even if an IRB decides such risks are outside its purview, ethical researchers should consider them. While well-informed research staff should be free to take on projects carrying non-zero risk (as much research entails), supervisors should aim to minimize the risk to research staffers, especially if they are students who may be reluctant to refuse. Designs such as passing through security checkpoints, for instance, may satisfy this requirement, at least if the researchers breach no laws (such as prohibitions of recording devices).

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. While what they measure is limited, it is exactly the thing that observational tools have in most contexts been unable to measure effectively: disparate treatment among similarly situated *individuals*, rather than neighborhoods.

## IV. CONCLUSION

Empirical research on race and policing poses many challenges, but it is worth trying to overcome these challenges because the stakes of the legal and policy debates such research seeks to inform are high. In many communities of color, intensive police presence fundamentally shapes daily life. Racial disparities in policing have recently come to the forefront of the national conversation, but they are not new; despite decades of research, we still do not have a clear picture of the reasons for them.

Constitutional litigation can be a valuable tool for redressing disparities, and constitutional doctrine specifically asks us to identify whether racially disparate outcomes are the result of disparate treatment by the police. Many police departments themselves care about this question, having committed to the elimination of racial profiling. But in most contexts, we simply do not have the data and the statistical tools to engage in this kind of causal analysis, and we may need to turn to new ways of generating new kinds of data that allow

---

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

[185] 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 131, at 126.

more rigorous analyses. The use of testers is one approach worth considering seriously.

There are multiple promising empirical strategies for analyzing racial disparities, and I do not suggest that the use of field experiments is likely to displace the need for careful observational analyses. Such analyses have already provided useful insights on some questions, and I have suggested some ways to push observational research further, such as the creation of more ambitious surveys about behavior and police-citizen contacts. But such research will always face omitted-variable and causal-inference challenges, and experimental work can be a very useful supplement. Current research has, in substantial ways, fallen short, despite decades of serious and resource-intensive efforts. It is time to think creatively about new solutions.

