Reconciling Legal and Empirical Conceptions of Disparate Impact: An Analysis of Police Stops Across California

Joshua Grossman*
Stanford University, Stanford, California, USA

Julian Nyarko
Stanford Law School, Stanford, California, USA

Sharad Goel
Harvard University, Cambridge, Massachusetts, USA

Abstract

We evaluate the statistical and conceptual foundations of empirical tests for disparate impact. We begin by considering a recent, popular proposal in the economics literature that seeks to assess disparate impact via a comparison of error rates for the majority and the minority group. Building on past work, we show that this approach suffers from what is colloquially known as "the problem of inframarginality", in turn putting it in direct conflict with legal understandings of discrimination. We then analyze two alternative proposals that quantify disparate impact either in terms of risk-adjusted disparities or by comparing existing disparities to those under a statistically optimized decision policy. Both approaches have differing, context-specific strengths and weaknesses, and we discuss how they relate to the individual elements in the legal test for disparate impact. We then turn towards assessing disparate impact of search decisions among approximately 1.5 million police stops recorded across California in 2022 pursuant to its Racial Identity and Profiling Act (RIPA). The results are suggestive of disparate impact against Black and Hispanic drivers for several large law enforcement agencies. We further propose alternative search strategies that more efficiently recover contraband while also exerting fewer racial disparities.

Keywords: discrimination, disparate impact, policing

1 Introduction

Anti-discrimination law in the U.S. recognizes two distinct forms of discriminatory conduct. First, disparate treatment law aims at prohibiting decisions that intentionally condition on protected group status, either to harm minorities or as an intermediate step to further a different goal. This notion of discriminatory conduct is largely consistent with Becker's popular model of discrimination, which defines as discriminatory those actions that are

*Corresponding author. Department of Management Science and Engineering, Stanford University, 475 Via Ortega, Stanford, CA 94305, USA. Email: jdgg@stanford.edu

either motivated by animus ("taste-based discrimination") or that use race as a proxy for an unobservable, decision-relevant factor ("statistical discrimination") [Becker, 1957]. But in many areas of life, such as employment and credit, U.S. law has long embraced a second definition of discrimination, significantly broadening its scope. This notion of discrimination is known as disparate *impact*. The doctrine of disparate impact renders illegal those policies that produce avoidable and unjustified excess disparities [Griggs v. Duke Power Co., 401 U.S. 424 (1971)]. In practice, disparate impact is often found if the plaintiff can demonstrate the existence of an alternative, feasible decision rule that is at least as good as the existing decision rule at achieving the stated policy goal while imposing fewer disparities. Although the law's embrace of disparate impact doctrine can be traced back decades, empirical scholarship both in the law and in the social sciences at large has almost exclusively limited itself to the analysis of disparate treatment. It is only recently that the literature has made a systematic attempt to broaden its focus [Arnold et al., 2022, Ayres, 2005, 2010, Bartlett et al., 2021, Bohren et al., 2022, Cai et al., 2022, Elzayn et al., 2023, Grossman et al., 2024, Grunwald et al., 2022, Jung et al., 2023]. In its wake, several statistical frameworks for the measurement of disparate impact have been proposed independently.

In this paper, we introduce, compare and critically assess the three most prominent recent proposals: risk-adjusted disparities, disparities relative to statistically optimized decision policies, and error-rate disparities. We discuss their individual advantages and disadvantages, and examine how they relate to the legal doctrine of disparate impact as it has been developed by U.S. courts. We argue that the first two proposals speak to different legal elements of disparate impact doctrine, making both valuable empirical tools for assessing disparate impact in various situations, albeit with context-specific strengths and weaknesses. However, we further argue that the third proposal—error-rate disparities—is generally unsuitable for assessing disparate impact. To foreshadow our argument, consider a judge who makes detention decisions by balancing public safety with the individual rights of the defendants. Imagine the judge is able to perfectly distinguish between "risky" and "non-risky" defendants, and chooses to only detain risky defendants. Assuming further that detention decisions were generally subject to disparate impact law,* the judge's decision practice would nonetheless not be considered to exert a disparate impact. After all, the judge is making decisions that optimally fulfill the goal of balancing public safety with the defendant's interests, and any residual disparities that result would thus be considered justified. Yet, as we show below, in most scenarios a measure that relies on error rates would find that the judge's decision practice is illegal, a result that can be reconciled neither with disparate impact doctrine nor with existing normative notions of discrimination. At a technical level, we illustrate that such error-rate-based measures suffer from what is known as the "problem of inframarginality" [Ayres, 2002, Hedden, 2021, Simoiu et al., 2017], an observation previously made in the technical literature in the context of algorithmic decisions [Corbett-Davies et al., 2017, 2023].

We utilize the insights obtained from our discussion and apply them to the concrete example of search decisions during vehicle and pedestrian stops conducted by police officers. To that end, we analyze a novel dataset of approximately 1.5 million stops recorded by

*Disparate impact is only illegal if a statute deems it so.

police agencies across California in 2022. During this time period, officers could choose to search stopped pedestrians and drivers whom they perceived as sufficiently likely to be carrying contraband. Officers were, on average, more likely to search Black and Hispanic individuals than white individuals, providing *prima facie* evidence of disparate impact. Moving beyond this *prima facie* evidence, we then apply the proposals above to assess the evidence for disparate impact. Following the first proposal, we compute risk-adjusted disparities to determine whether the gap in search rates is justified by legitimate policy goals—namely, the recovery of contraband. To do so, we estimate the statistical likelihood that a stopped individual is carrying contraband, using all available recorded information. We find that search rates for stopped Black and Hispanic individuals are considerably larger than for stopped white individuals of comparable risk (i.e., the racial disparities persist even after accounting for risk). Next, following the second proposal, we compare the observed racial disparities to those achievable under a set of statistically optimized alternative search policies. We specifically consider a set of "threshold" policies, in which all individuals above a given level of estimated risk are searched. We find that there are indeed alternative policies that: (1) recover more contraband than the status quo; (2) require conducting fewer searches; and (3) impose fewer racial disparities. The existence of such policies provides additional evidence of disparate impact. We emphasize, however, that these results are primarily intended as an illustration of preferred approaches to measuring disparate impact. To conclusively demonstrate illegal disparate impact in a court setting, further scrutiny of policing practices is necessary.

Embracing a broader concept of anti-discrimination is vital in ensuring that empirical scholarship remains closely tied to legal realities. Our hope is that this study can contribute to that goal by serving as a guide to researchers interested in assessing the disparate impact of a policy or decision rule.

2 The Law of Disparate Treatment and Disparate Impact

Although our focus lies on disparate impact law, a brief primer on the legal concepts in U.S. anti-discrimination can serve as helpful background. For ease of exposition, we will focus on the law surrounding racial discrimination, although most of the content equally applies to other forms of discrimination based on legally protected features, such as gender.

Generally speaking, U.S. law recognizes two forms of discriminatory conduct: disparate treatment and disparate impact. Disparate treatment encapsulates the most intuitive notion of discrimination. It is aimed at outlawing decisions and policies that are motivated by race, making discriminatory intent the crucial element of disparate treatment [DeJung v. Superior Ct., 169 Cal. App. 4th 533 (2008); McDonnell Douglas Corp. v. Green, 411 U.S. 792 (1973)]. The intent can take the form of explicit, racially conditioned decision making. But more commonly, disputes focus on facially neutral decisions or policies that are alleged to be—at least in part—racially motivated. Disparate treatment by public entities is governed by the Equal Protection Clause of the U.S. Constitution. Accordingly, if discriminatory intent is present and the discriminatory actor is a public entity or official, the decision is subjected to judicial review under a "strict scrutiny" standard [United States

v. Carolene Prod. Co., 304 U.S. 144 (1938)].* This standard is very difficult to meet and requires that the conduct in question is *narrowly tailored* to serve a *compelling state interest*. The only examples relevant today in which race-based decisions met this standard consist of affirmative action cases in a handful of domains, such as in government contracting [Rothe Dev., Inc. v. United States Dep't of Def., 836 F.3d 57 (D.C. Cir. 2016)] and—until recently—education [Fisher v. Univ. of Texas at Austin, 579 U.S. 365 (2016); Students for Fair Admissions, Inc. v. President & Fellows of Harvard Coll., 143 S. Ct. 2141 (2023)]. In addition to the constitutional constraints imposed on public actors, disparate treatment by private actors is outlawed by federal and state laws in many public-facing contexts,† although relevant nuances may vary by context.‡

Although constitutional and statutory prohibitions against disparate treatment developed separately and employ differing evidentiary standards [Harris, 2014], they share a strong emphasis on the element of discriminatory *intent*. It is thus common to conceptualize disparate treatment as decisions made *because of* race.§ Empirical legal research has translated this definition of disparate treatment into a "kitchen-sink" model [Gaebler et al., 2022]. Under this approach, the investigator typically runs a regression of the form

$$\Pr(Y_i = 1) = \text{logit}^{-1}\left(\alpha_{\texttt{race}[i]} + \beta^T \vec{X}_i\right), \tag{1}$$

with Y_i indicating a binary decision, $\alpha_{\texttt{race}[i]}$ an intercept term shared by defendants with the same race or ethnicity as defendant i, and $\vec{X}_i$ a vector of additional controls. The controls included in $\vec{X}$ are typically expansive. The idea behind the kitchen-sink approach is that $\vec{X}$ controls for non-racial factors that might motivate the decision (e.g., to search a stopped individual). Thus, any residual variation that is explained by $\alpha_{\texttt{race}}$ holding the covariates in $\vec{X}$ constant is taken as evidence of discriminatory intent. Following this logic, Wooldredge [2012] seeks to provide evidence for disparate treatment of Black pretrial defendants by fitting models of the form above that estimate detention rates after adjusting for legally relevant factors, including (non-racial) demographics, prior criminal history, and charges. Many other studies, especially in criminal law, follow a similar process [Bridges and Steen, 1998, Demuth, 2003, Didwania, 2020, Donnelly and MacDonald, 2018, Metcalfe and Chiricos, 2018, Rehavi and Starr, 2014].

*The now widely disparaged case of *Korematsu*, although now overturned, is also among the cases that contributed to the development of the strict scrutiny standard [Korematsu v. United States, 323 U.S. 214 (1944)].

†A notable exception is insurance, where disparate treatment is not outlawed in all states [Avraham et al., 2013].

‡For instance, 29 C.F.R. § 1608 lays out detailed guidelines under which voluntary affirmative action efforts by private employers are protected.

§There still is substantial disagreement surrounding the details of discriminatory intent. For the constitutional context, see Huq [2017]. For the context of Title VII, see Oppenheimer [1992] and Bornstein [2017]. In addition, the case law has since developed to also include other notions of disparate treatment. For instance, even if a policy was originally instituted with good intentions, under the concept of "deliberate indifference", some courts have found it can still constitute disparate treatment if the policy's disproportionate, negative impact on minorities is known and the policy is not corrected within a reasonable time frame [Davis Next Friend LaShonda D. v. Monroe Cnty. Bd. of Educ., 526 U.S. 629 (1999); Floyd v. City of New York, 959 F. Supp. 2d 540 (S.D.N.Y. 2013)].

There are some problems with conceiving of disparate treatment in this way. Among others, it is our view that empirical researchers often define the set of covariates included in $\vec{X}$ too broadly. Because every variable in $\vec{X}$ is implicitly accepted as being free of racial motivation, being too broad can quickly lead researchers to mask discriminatory intent if the discriminatory practice is implemented through a facially neutral factor.*. However, a full discussion of statistical measures of disparate treatment is beyond the scope of this paper.

In addition to disparate treatment, U.S. anti-discrimination laws sometimes render illegal a second form of discriminatory conduct, disparate impact. But unlike disparate treatment, there is no general prohibition of disparate impact under the U.S. Constitution. Instead, disparate impact is rendered illegal only through state and federal laws. The most prominent domains subject to disparate impact analysis include credit [15 U.S.C. § 1691 et seq.], employment [42 U.S.C. § 2000e et seq.] and housing [42 U.S.C. § 3601 et seq.]. Although the fragmented nature requires a few generalizations, disparate impact laws aim to prevent policies and decisions that, while not necessarily racially motivated, nonetheless have an adverse impact on racial minorities that cannot be justified by a furtherance of the policy goals.

To illustrate, consider the case of a job posting by a tech company for the position of a software engineer. The posting requires applicants to have a computer science degree. The degree requirement impacts Black potential applicants more negatively than white potential applicants, given that the share of Black computer scientists is disproportionately low [Dillon Jr et al., 2015]. However, a computer science degree can reasonably be assumed to teach skills that software engineers benefit from, meaning that the degree requirement does not constitute disparate impact. But contrast this to the seminal case of *Griggs v. Duke Power Co.*, where the Supreme Court examined an internal policy under which a high school diploma was required for certain promotions within Duke Power Company in North Carolina. Black employees were much less likely to hold a high school diploma than white employees, thus disproportionately excluding the Black minority from the positions. The Supreme Court found that, while it is principally permitted to impose job requirements that impact racial minorities disproportionately, a high school diploma did not indicate better job performance, thus rendering the requirement illegal.

More formally, legal tests of disparate impact typically have three elements. Those require that: (1) the minority group is disproportionately impacted by a policy ("adverse impact") [New York City Env't Just. All. v. Giuliani, 214 F.3d 65 (2d Cir. 2000)]; (2) that there is no legitimate justification for the policy [Texas Dep't of Hous. & Cmty. Affs. v. Inclusive Communities Project, Inc., 576 U.S. 519 (2015)]; and (3) that an alternative policy with a lesser disproportionate impact is available and implementable [Elston v. Talladega Cnty. Bd. of Educ., 997 F.2d 1394 (11th Cir. 1993)]. The plaintiff is responsible for establishing that the defendant's policy adversely impacts the minority group. The burden then shifts to the defendant, who must show that the adverse impact is justified by legitimate policy goals. Failing to do so would typically result in a finding of disparate impact. However, if the defendant does provide compelling justification for the disparities, the burden then shifts back to the plaintiff, who, to establish a finding of disparate impact,

*We note that there are additional problems with defining the causal effect of race on decisions [Gaebler et al., 2022, Sen and Wasow, 2016].

must show that there exists an equally efficient policy with less adverse impact than the status quo [see, e.g., 42 U.S. Code § 2000e–2]. We describe these three elements in more detail below.

Adverse Impact An adverse impact is typically defined as the difference in group-based selection rates [29 C.F.R § 1607.16]. In the context of standardized tests for promotions, for instance, a court would compare the passage rate among white test takers to the passage rate among Black test takers. The test would demonstrate an adverse impact if the passage rate among Black test takers was substantially lower than that among white test takers.*

No Justification An adverse impact lacks a substantial justification if it is not demonstrably related to a significant, legitimate goal. At times, it is also held that the adverse impact needs to be a *necessary condition* to effectuate the policy goal.† How courts operationalize this requirement is highly context-specific [Clady v. Los Angeles Cnty., 770 F.2d 1421 (9th Cir. 1985); Smith v. Xerox Corp., 196 F.3d 358, 363 (2d Cir. 1999); Groves v. Alabama State Bd. of Educ., 776 F. Supp. 1518 (M.D. Ala. 1991)]. For instance, the strength of the evidence required may vary by the extent of the adverse impact, by the entity that makes the relevant decision, and by whether the decision-relevant factors that cause the disparity are innate or can be acquired.

No Less Discriminatory Alternative Demonstrating the shortcomings of the current policy is not enough if there is no less discriminatory alternative [Elston v. Talladega Cnty. Bd. of Educ., 997 F.2d 1394 (11th Cir. 1993); Georgia State Conf. of Branches of NAACP v. State of Ga., 775 F.2d 1403 (11th Cir. 1985)]. In this way, disparate impact law is grounded within the realm of feasible policy choices: If the only way for an employer to mitigate adverse impact is to spend tens of thousands of dollars on each applicant to assess their suitability for the job, this is not something that anti-discrimination laws will ask of them. With the advent of algorithmic decision making, the requirement to have no less discriminatory alternative has received heightened relevance. Often, if a decision was based on these complex model estimates, it would both improve outcomes and decrease the adverse impact [Goel et al., 2016]. However, it remains unclear in what contexts decision makers will be required to rely on these more complex estimation procedures. Does disparate impact law require employers to forego their traditional, interview-based hiring practices if it can be shown that algorithmic assessments of job performance are superior and impose fewer disparities [Hoffman et al., 2018]? To date, courts have shied away from providing a clear answer.

*In the employment context, courts often apply a four-fifths rule, under which the difference is consequential if the passage rate for Black test takers is less than 80% of the passage rate of white test takers [29 C.F.R. § 1607.4].

†In which case the dividing line between the justification requirement and the requirement for a less discriminatory alternative is blurred.

3 Statistical formulations of disparate impact

Unlike for disparate treatment, there have been surprisingly few attempts to provide a statistical framework for the evaluation of disparate impact. Our goal in this section is to introduce and mediate between the different approaches. We focus on three statistical formulations, all of which are relatively recent.

3.1 Differences in error rates

One approach to measuring disparate impact is rooted in error rates. This approach deems discriminatory those decisions that lead to differences in error rates across the marginalized and the majority group, such as the false positive or the false negative rate. Conceiving of biases as error rates has a long tradition in the literature on algorithmic fairness in computer science and statistics [Buolamwini and Gebru, 2018, Chouldechova, 2017, Corbett-Davies et al., 2017, Dwork et al., 2012, Kleinberg et al., 2017], law [Chander, 2016, Huq, 2019, Mayson, 2019], medicine [Goodman et al., 2018, McCradden et al., 2020, Paulus and Kent, 2020], the social sciences [Berk et al., 2021, Imai et al., 2023, Kleinberg et al., 2018], and philosophy [Card and Smith, 2020, Hu and Kohler-Hausmann, 2020, Kasy and Abebe, 2021]. This formulation of bias has not typically been tied to disparate impact law. But a recent contribution in the economics literature has proposed a measure based on error rates that is explicitly described as an estimand corresponding to the legal concept of disparate impact [Arnold et al., 2021, 2022, Baron et al., 2023]. This estimand—and its associated, novel estimation method—have since attracted significant attention.

Arnold et al. [2022] illustrate their measure of disparate impact in a pretrial detention setting in which judges must decide whether or not to release defendants on bail. Each defendant has a latent "misconduct potential", which takes on the value 1 if the defendant would violate the terms of release if released, and 0 if not. Their measure of disparate impact, Δ, is based on a weighted sum of the difference in true negative rates and the difference in false negative rates across two groups of individuals, with weights defined by the overall violation rate across all individuals. Arnold et al. [2022] use the following mathematical formulation:

$$\Delta = (\delta_w^T - \delta_b^T)(1 - \bar{\mu}) + (\delta_w^F - \delta_b^F)\bar{\mu}, \tag{2}$$

where δ_r^T is the true negative rate for individuals of race r (i.e., the proportion released among those who would not violate if released), δ_r^F the false negative rate for individuals of race r (i.e., the proportion released among those who *would* violate if released), and $\bar{\mu}$ the expected violation rate if all individuals were released. Here, for exposition, w refers to white defendants, and b refers to Black defendants.

Drawing on past work in computer science [Corbett-Davies et al., 2023], we argue that any such measure of disparate impact (or "fairness," for that matter) that is based on error rates is ill-suited to provide either legal or policy guidance. This is because these measures suffer from what is colloquially known as the "problem of inframarginality" [Ayres, 2002]. Intuitively, the problem is that error rates do not only capture aspects of the decision rule, but also of the underlying risk distribution for each group [Simoiu et al., 2017]. When defining disparate impact in such a way, an actor who does the best possible job to make

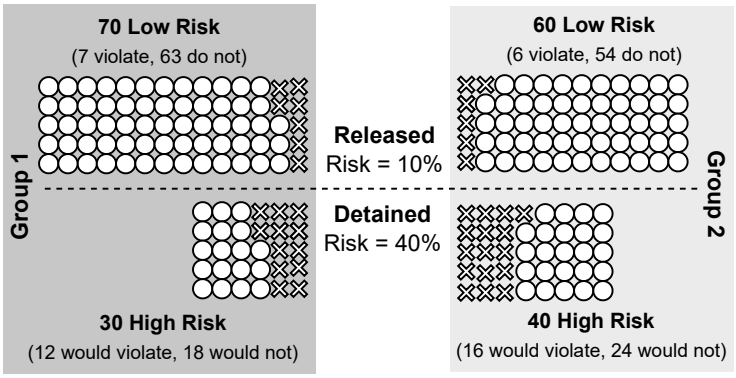

Figure 1: *Illustration of the problem of inframarginality when comparing error rates across groups with different underlying distributions of risk. Suppose there are two groups of pretrial defendants and two possible levels of pretrial risk. Each group has 100 defendants. If released pretrial, lower-risk defendants violate the terms of release 10% of the time, and higher-risk defendants violate 40% of the time. 30% of Group 1 defendants are higher risk, compared to 40% of Group 2 defendants. Suppose a judge can perfectly perceive pretrial risk. The judge imposes a unilateral risk threshold decision rule in deciding whom to release: lower-risk defendants are always released, and higher-risk defendants are always detained. In this scenario, the* Δ *measure of disparity from Arnold et al. [2022] is approximately* 0.1*, incorrectly suggesting that the decision rule disparately impacts Group 2 defendants. See main text for calculations.*

decisions in furtherance of the stated policy goal can be found to discriminate simply due to differences in underlying risk distributions. In an attempt to avoid liability for disparate impact under this definition, the actor would then be required to make contra-indicated decisions, such as to search or jail people that, to the best of their knowledge, are of low risk.

We illustrate this argument by way of a specific example in the context using the Δ estimand proposed by Arnold et al. [2022], shown in Figure 1. For a more formal treatment that extends to a wider range of error-rate-based measures and also discusses the case of different thresholds for each group, see Corbett-Davies et al. [2023]. Suppose there are two groups of pretrial defendants, each with 100 defendants. Each defendant has either a 10% likelihood of violating the terms of pretrial release if released ("low risk") or a 40% likelihood ("high risk").* Imagine 30% of defendants in Group 1 are of high risk, and 40% of defendants in Group 2 are of high risk. Further suppose that the presiding judge can perfectly estimate whether a defendant is of low risk or high risk, using only legally permissible factors. The judge decides whether to detain defendants based on a simple rule: high-risk defendants are detained, and low-risk defendants are released. Denote a true negative as an instance in which the judge releases a low-risk defendant, and denote a false negative as an instance in which the judge releases a high-risk defendant.

*For simplicity, we use groups of equal size with only two possible risk levels. This particular example is amenable to groups of different size, with two unique risk levels for each group. In Appendix Figure A1, we extend the example to a setting with a continuous distribution of risk.

Although this decision rule treats similarly situated* defendants identically, the true negative rates and false negative rates among each group differ in expectation. In this example, the true negative rate is the proportion who are released, among those who *would not* violate if released. Here, 81 of the defendants in Group 1 would not violate if released (as indicated by the circles in the left-hand side of Figure 1). Further, 63 of these defendants are actually released—represented by the ◯ symbols above the dotted line—resulting in a true negative rate of $63/81 = 78\%$. We can analogously compute the true negative rate for Group 2. In particular, among the 78 defendants from Group 2 who would not violate if released (the ◯ symbols on the right-hand side of Figure 1), 54 are released (those above the dotted line), yielding a true negative rate of $54/78 = 69\%$. Importantly, the true negative rates differ across groups even though the same, risk-conditioned decision rule was applied to each group.

Similarly, the false negative rate is the proportion who are released, among the defendants who *would* violate if released. In Group 1, 19 defendants would violate if released (represented by the × symbols on the left-hand side of Figure 1). Among these defendants, 7 are released (the × symbols above the dashed line), resulting in a false negative rate of $7/19 = 37\%$. Moving to Group 2, 22 defendants would violate if released (indicated by the × symbols on the right-hand side of Figure 1). Among these 22 defendants, 6 are released (those above the dashed line), giving us a false negative rate of $6/22 = 27\%$.

Next, the calculation of Δ requires computing $\bar{\mu}$, which is the expected violation rate that would result from releasing all 200 defendants. Among the 200 defendants depicted in Figure 1, there are 41 defendants who would violate if released (represented by the × symbols), yielding an overall violation rate of $41/200 = 21\%$. Finally, we compute Δ using the results above:

$$\begin{aligned}\Delta &= (\delta_1^T - \delta_2^T)(1-\bar{\mu}) + (\delta_1^F - \delta_2^F)\bar{\mu} \\ &= (0.78 - 0.69)(1 - 0.21) + (0.37 - 0.27)(0.21) = 0.1.\end{aligned}$$

The resulting value of $\Delta = 0.1$ suggests disparate impact to the disadvantage of defendants in Group 2. The only way for the judge to reduce Δ is to detain some low-risk defendants and/or release some high-risk defendants.

Figure 2 extends the example in Figure 1 to a range of similar scenarios in which a judge only detains high-risk defendants. The scenario in Figure 1, in which 30% of Group 1 defendants and 40% of Group 2 defendants are high risk, is denoted by the × symbol in the fourth panel of Figure 2. Appendix Figure A1 further extends this example to continuous distributions of risk.

*Where similarly situated is with respect to the goal of the release policy, which is to release as many defendants as possible while minimizing pretrial violations.

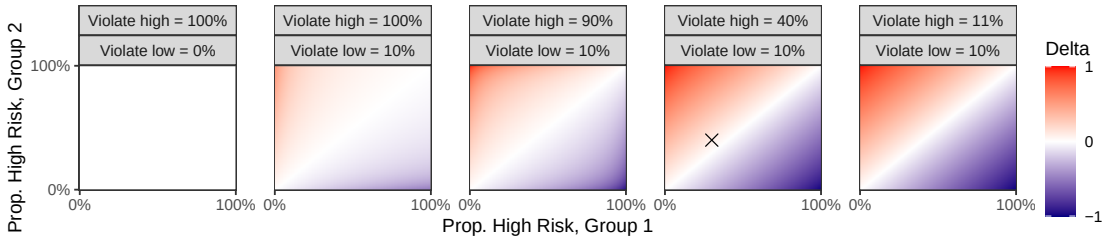

Figure 2: *Extension of the scenario in Figure 1 to different discrete distributions of risk and different violation probabilities. The* $\times$ *symbol panel denotes the scenario from Figure 1. When risk distributions are identical, the* Δ *(Delta) measure of disparity correctly indicates no disparate impact, as indicated by the white diagonal in each panel. The leftmost panel shows that the* Δ *measure correctly indicates no disparate impact when low-risk defendants never violate and high-risk defendants always violate. However, as the violation probabilities of low- and high-risk defendants move away from the extremes, differences in the underlying distributions of risk result in non-zero values of* Δ*, incorrectly indicating evidence of disparate impact.*

Overall, these results illustrate the problem of infra-marginality that plagues error rates: Error rates change as a function of the underlying group-specific risk distributions, even if the decision rule remains the same. Hence, in these simulations, the Δ measure correctly indicates the absence of disparate impact in only two scenarios: (i) when risk is perfectly predictive, such that high risk defendants *always* carry contraband and low-risk defendants *never* do; or (ii) when the risk distributions between the groups are identical. In all other cases, the Δ measure indicates disparate impact on Group 2 if Group 2 defendants are, on average, riskier than Group 1 defendants, and vice versa. As the violation probabilities for low-risk and high-risk defendants move from the extremes, the Δ measure of disparate impact is more sensitive to differences in the distribution of risk across groups. Because the measure does not allow us to draw accurate inferences about the decision rule, we believe it is ill-suited to capture disparate impact. Indeed, it is our view that this and other estimands based on error rates are not appropriate to accurately capture notions of fairness, calling into question their utility [Chohlas-Wood et al., 2023, Corbett-Davies et al., 2023, Hedden, 2021, Simoiu et al., 2017].

Alternative Interpretations of Error Rate Differences

One interpretation of error rates is that they are a direct measure of disparate impact [Arnold et al., 2022]. An alternative understanding of error rate discrepancies is that they merely provide a first signal that is neither necessary nor sufficient to establish disparate impact, but should give reason for further investigation. For instance, despite inframarginality issues, Hellman [2020] suggests that "[a] lack of error ratio parity between a previously disadvantaged group and its counterpart (blacks and whites, for example) is suggestive of unfairness and provides a normative reason to engage in further investigation and for caution." [Hellman, 2020, 845-846]. One could think of this conceptualization of error rate discrepancies as replacing the first element in a disparate impact analysis, thereby shifting the burden

of proof to justify the disparities towards the decision maker. If that decision maker then showed that the discrepancies arise from differences in the underlying risk distribution, they would be justified.

Although not subject to the same statistical problems, we similarly believe this conceptualization of error rates not to be fruitful. "Adverse impact" as the burden-shifting element, although not free of significant limitations, has the virtue of at least being simple to understand and easy to form statistically accurate intuitions about. For error rates, this is not so. Imagine an audit that finds detention rates among Black defendants to be 10% higher than for white defendants. It is quite intuitive to determine that this discrepancy alone is insufficient to establish the existence of a discriminatory decision making process [Ayres, 2010, Heaton et al., 2017]. This is because it is easily understood how risk-relevant differences between the groups may cause the identified discrepancies. However, if that same audit found the error rates for Black defendants to be 10% higher than for white defendants, those without empirical training would have trouble understanding whether or not this is definitive evidence for discrimination. After all, appreciating the problem of inframarginality requires at least some familiarity with statistics, which is often lacking among the relevant legal actors. As such, there is a concern that shifting from easily understood statistics like selection rates to more complex statistics like error rates could convolute the assessment. And because neither selection nor error rate differences alone provide strong evidence for disparate impact, we don't believe that shift would come with a tangible benefit that would justify the costs. In addition, differences in selection rates are a burden-shifting element that is firmly established in the relevant case law. Again, we see no reason to break with this significant precedent without a clear benefit on the other side. Last but not least, to the extent that error rates are not viewed as a *replacement* of adverse impact, but simply as an additional diagnostic tool to identify normatively problematic differences between two groups, we note that there are several more informative proxies. For instance, if it is believed that differences in risk distributions themselves point towards larger systematic discriminatory practices (e.g., overpolicing) [Hellman, 2020, 840], a simple solution would be to analyze these risk distributions directly, rather than to rely on error rates.

3.2 Risk-adjusted disparities

An alternative approach to the measure of disparate impact is risk-adjusted regression [Jung et al., 2023]. The approach consists of two steps. First, the analyst uses all available data to create a risk-model of the form

$$\texttt{risk}_i = g(X_i), \tag{3}$$

where $\texttt{risk}_i$ is the estimated risk of subject i, g is an arbitrary function, and $\vec{X}$ is a vector of included risk factors. Because the estimation of risk is purely predictive, g can, in principle, be arbitrarily flexible. For instance, risk can be estimated via a random forest or neural network. Similarly, X can contain an arbitrarily large set of covariates. In principle, one could even opt to include race itself (and other protected features) under the assumption that these covariates may capture risk-relevant but unobservable decision factors. At the

same time, we note that the inclusion of protected features could mask disparate treatment, especially in the form statistical discrimination, which could counsel against the inclusion.

In the second step, the analyst fits a model of the following (or similar) form:

$$\Pr(Y_i = 1) = \alpha_{\texttt{race}[i]} + \gamma \cdot \texttt{risk}_i, \tag{4}$$

where Y_i is the binary action taken by the decision maker, $\alpha_{\texttt{race}[i]}$ is an intercept term associated with the group of subjects with the same race or ethnicity as subject i, $\texttt{risk}_i$ is the estimated risk of subject i, as estimated from the first model, and γ is the coefficient of the risk term.*

Consider how this approach connects to the legal definition of disparate impact. Assuming that g is sufficiently flexible, the first model reflects the analyst's best attempt to capture an individual's probability that the relevant outcome (e.g., weapon recovery, recidivism or satisfactory job performance) will occur. If the actual decisions made were fully explainable by the individual's risk, the coefficient on $\alpha_{\texttt{race}}$ would be (close to) 0. But if instead the coefficient is significantly different from 0, this suggests that the actual decision rule imposes disparities that are not justified by risk. In this sense, a risk-adjusted regression speaks to the first two elements of a disparate impact claim. It can suggest the existence of an unjustified, adverse impact. At the same time, a risk-adjusted regression itself does not specify a specific, implementable policy, because it does not propose any particular decision rule. As such, it does not fulfill the third element, the showing of an alternative policy with less of an adverse impact.

3.3 Optimized decision making

In a scenario where risk or qualification can be estimated for every individual, the utility-maximizing decision rule is one where a unilateral threshold dictates decision making [cf. Corbett-Davies et al., 2023]. In other words, individuals with estimated risk or qualification above the threshold are selected, and individuals below are not. Among others, this approach has been used by Elzayn et al. [2023] to measure adverse impact under hypothetical risk thresholds. Similar to risk-adjusted regression, the first step consists of estimating a risk model of the form

$$\texttt{risk}_i = g(X_i). \tag{5}$$

After risk has been estimated, individuals are sorted based on their estimated risk. A threshold is drawn such that everyone above the threshold receives the costs/benefits and anyone below the threshold does not. After defining the threshold, adverse impact is assessed by comparing the group-specific probability of receiving the cost/benefit.

*The first-stage risk model aims to capture the "true risk" of an individual that could, in theory, be estimated by an observer at the scene immediately prior to the decision being made. To do so, we would typically include all available features, but that need not always be the case. For example, in the presence of label bias—where the outcome of interest is differentially recorded across groups—Zanger-Tishler et al. [2024] argue that the accuracy of a model can improve by strategically ignoring covariates. Similarly, if officers record risk factors in a biased manner, it is possible that ignoring information could yield a better estimate of risk-adjusted disparities relative to the true risk. To account for these possibilities, we estimate risk using all available information, and then, in the appendix, we compute the extent to which estimates of risk-adjusted disparities can vary as a function of how much our estimates of risk differ from the true risk.

How exactly the threshold is drawn is a matter of policy, and typically reflects some type of constraint. For instance, in defining a reference policy for the auditing practices of the IRS, Elzayn et al. [2023] pick the threshold such that the number of people audited are the same as under current IRS practices. Other possibilities are to draw a threshold such that the risk to public safety or the amount of loans given out are the same as under a current policy. For example, suppose a law enforcement agency seeks to assess potential disparate impact in its decisions to search stopped drivers. Using historical data, the agency estimates that they could have recovered the same amount of contraband had they searched all stopped drivers with a perceived risk of carrying contraband greater than 10%. Under this hypothetical policy, suppose that 15% of white drivers would have been searched, compared to 20% of stopped Black drivers (i.e., 1.33 times more often). Suppose that under the agency's actual policy, 25% of Black drivers were searched, compared to 10% of white drivers (i.e., 2.5 times more often). The existence of an implementable and equally-efficient policy with lower adverse impact suggests possible disparate impact in search practices.

Consider how this statistical approach relates to disparate impact law. Disparate impact law requires a showing of a feasible, alternative policy that has fewer disparities while achieving the stated policy goal at least as effectively as the current policy. If such a policy exists, it implies that the (greater) disparities under the current decision rule are avoidable. In this way, disparate impact law can be understood as a search over the policy space for policies that fulfill the before-mentioned criteria. This approach is equivalent to assessing a subset of the policy space for whether it provides less disparate alternatives. Importantly, threshold rules are not a random subset of decision rules. Instead, as Corbett-Davies et al. [2023] and others have shown, threshold decision rules are uniquely optimal among all policies, given estimated risk.

Both risk-adjusted regression and the search for risk-based alternative policies require an estimation of risk, reflected in g. As noted, due to the predictive nature of risk estimation, g can be arbitrarily flexible and can take an arbitrarily large set of covariates X as input. However, as detailed above, disparate impact law requires the plaintiff to propose alternative policies that are feasible and implementable. Depending on the context, it may be argued that such feasibility requires the imposition of constraints, both on g and on X. Take, for instance, the decision to search a stopped pedestrian, which is often a split-second choice that is made in the moment. If g takes a complex functional form such as a neural network, the model will uncover statistical associations that a police officer who is patrolling the beat might not be able to uncover themselves. Thus, the only way for the officer to meet the standard implicitly set by the use of g would be for them to use the risk model themself, e.g., by feeding a feature vector for the potential suspect into the model and obtaining the prediction. This is not always realistic, and so we may want to confine g to resemble decision making rules that the officer can quickly employ while on patrol. Such concerns are of less relevance, however, if well-resourced actors are making decisions without imminent time constraints. Indeed, some entities are already using complex algorithms, as is the case when the IRS makes its auditing decisions [Elzayn et al., 2023]. Allowing g to be flexible in such contexts is merely akin to a requirement that they use the best available algorithm, which can often be achieved with relative ease. In the next section, we discuss the implementability of threshold rules in more detail.*

*We also note that optimized decision making can still be fruitfully implemented even though estimated

4 Measuring disparate impact in policing

To illustrate the discussed approaches, we next focus on a novel dataset of approximately 1.5 million stops to estimate disparate impact in search decisions of law enforcement agencies across California. In doing so, we highlight that the liability of law enforcement agencies under existing disparate impact laws is, as a legal matter, highly theoretical and contested [Tiwari, 2019]. However, legal irrelevance does not imply policy or normative irrelevance, and we believe that disparate impact analysis can contribute significantly to better policymaking, irrespective of the specific, statutory context. This is for at least two reasons: First, a showing of disparate impact entails the proposal of an alternative, equally efficient yet less disparate policy. As such, it provides a concrete and practical way to make better policy. Second, and relatedly, a finding of unjustified disparities can exert pressure to examine an existing, presumably harmful policy.

The California State Legislature passed the Racial Identity and Profiling Act (RIPA) in 2015. RIPA requires that officers record detailed information following every stop of a pedestrian or driver, including information on race, ethnicity, and other protected characteristics.* We limit our analysis to 1,604,926 stops conducted in 2022 by the 50 agencies in California recording the most stops that year.† Among this initial set of stops, we exclude those for which the officers often conduct non-discretionary searches. Specifically, we exclude stops for which the stated stop reason was either (1) knowledge of parole/probation/postrelease community supervision (PRCS)/mandatory supervision or (2) knowledge of arrest warrant.‡ We additionally exclude consensual encounters, since only a subset of these interactions are recorded in our data, namely those that resulted in a search. Finally, we exclude stops initiated for suspected truancy or other educational policy, as the threshold for conducting searches in an educational context can often be lower [People v. William G. (1985) 40 Cal.3d 550]. This filtering results in 1,516,316 stops for our primary analysis. Next, for this set of stops, we consider a "search" to have taken place if, and only if, it appears that officers exercised discretion in determining whether to search the individual. In particular, we do not consider a stop to have resulted in a "search" if the exclusive recorded reasons for the search were limited to: (1) search warrant, (2) condition of parole/probation/PRCS/mandatory supervision, (3) incident to arrest, or (4) vehicle inventory.

Table 1 displays summary statistics for these 1.5 million stops, disaggregated by race and ethnicity. Overall, 45% of stopped individuals were Hispanic, 29% were white, 17% were Black, and 10% were of another race or ethnicity.§ Across all groups, moving violations were the most common reason for conducting a stop, though Black and Hispanic individuals were

risk scores are biased, e.g. due to disparate recording practices. In effect, the goal of such an analysis then is to demonstrate that disparities can be reduced via alternative policies, taking the (biased) risk scores as given.

*Figures A2, A3, and A4 are excerpts from the RIPA data collection form.

†We exclude from the RIPA data stops conducted by seven large agencies for which we cannot accurately assess risk. See Figure A6 for details.

‡Figure A5 shows that the search and arrest rates for these stop reasons are unusually high relative to other reasons.

§Approximately 99% of individuals in the data are recorded as identifying with a single race or ethnicity. The remaining 1% are considered 'Hispanic' if they are classified as 'Hispanic' along with any other race or ethnicity, 'Black' if they are classified as 'Black' and any other race or ethnicity other than Hispanic, and 'Other' otherwise.

more likely to be stopped for an equipment violation than individuals from other groups, and Black individuals were the most likely to be stopped for suspected criminal activity. 20% of stopped Black individuals and 15% of stopped Hispanic individuals were ostensibly searched at the discretion of the officer, compared to 9% of stopped white individuals and 5% of stopped individuals from other groups. The difference in search rates between stopped Black and Hispanic individuals and stopped white individuals could serve as the *prima facie* adverse impact component of a disparate impact claim.*

Table A1 separately reports adverse impact for the California law enforcement agencies with the most stops recorded in 2022 (henceforth the "largest" agencies). For many agencies, Black and Hispanic individuals were searched at substantially higher rates than white individuals. For expositional purposes, we focus the main analysis on the 10 largest agencies: the Los Angeles Police Department, the Los Angeles County Sheriff, the San Diego Police Department, the San Bernardino County Sheriff, the Riverside County Sheriff, the San Jose Police Department, the Orange County Sheriff, the Sacramento Police Department, the Sacramento County Sheriff, and the Anaheim Police Department. We include expanded results for the 50 largest agencies in the Appendix.

4.1 Application of risk-adjusted regression

To determine whether the observed adverse impact in search rates across agencies may be justified, we first measure risk-adjusted disparities in search rates. Then, we attempt to construct alternative search policies with lower adverse impact and the same or greater efficiency than the status quo search policies. Under both approaches, we find suggestive evidence that search practices in many California law enforcement agencies may have imposed a disparate impact on Black and Hispanic individuals. As emphasized in the introduction, these results are not conclusive evidence of illegal disparate impact within particular California law enforcement agencies.†

*Here we focus on disparate impact in search decisions conditional on being stopped, not stop decisions themselves. Although a comprehensive analysis of disparities should also assess the latter, we do not have data on those who were not stopped, making it impossible to estimate the associated disparities absent additional assumptions. We do, however, present the results of assessing adverse impact using the local demographic distribution in Table A1.

†For instance, individual-level results could be sensitive to internal agency policies that are unobservable to us at scale.

Variable	All	Black	Hispanic	Other	White
Num. stops	1,516,316	252,956	686,367	142,179	434,814
Prop. all stops	1.00	0.17	0.45	0.09	0.29
Equipment viol.	0.19	0.21	0.21	0.16	0.16
Moving viol.	0.46	0.37	0.46	0.63	0.47
Non-moving viol.	0.10	0.11	0.09	0.08	0.10
Sus. crim. activity	0.25	0.31	0.24	0.13	0.26
Searched (any)	0.20	0.30	0.22	0.09	0.15
Searched (discretion)	0.13	0.20	0.15	0.05	0.09
Hit rate (any)	0.28	0.28	0.27	0.28	0.30
Hit rate (discretion)	0.30	0.32	0.29	0.30	0.34
Non-discretionary	0.35	0.34	0.32	0.41	0.44
Consent	0.28	0.21	0.32	0.27	0.27
Safety	0.33	0.36	0.35	0.30	0.27
Suspect weapon	0.11	0.15	0.11	0.08	0.06
Plain view	0.08	0.12	0.07	0.06	0.07
Evidence of crime	0.08	0.08	0.07	0.08	0.07
Plain smell	0.04	0.08	0.04	0.02	0.01
Emergency	0.005	0.01	0.004	0.01	0.004
Canine	0.002	0.001	0.002	0.002	0.002

Table 1: *Summary statistics for the 1.5 million RIPA stops included in the analysis. The first block of statistics shows racial demographics and stop reasons. Most individuals were stopped because of a moving violation. Black individuals were more likely to be stopped for suspected criminal activity than all other groups, while Black and Hispanic individuals were the most likely to be stopped for an equipment violation. The second block shows the proportion of stopped individuals who were searched, and the proportion of searches that resulted in contraband recovery ("hit rate"), with separate statistics for discretionary searches. Stopped Black and Hispanic individuals were more likely to be searched than stopped white individuals, and stops of white individuals were the most likely to result in a contraband recovery. The final block shows the prevalence of each recorded search basis among stopped individuals who were also searched. The most common search bases were non-discretionary, such as an outstanding arrest warrant or a condition of parole. The most common discretionary bases were consent of the stopped individual and officer concern for their own safety or the safety of others.*

To generate risk estimates required for both risk-adjusted regression and risk-thresholded decision rules, we fit, separately for each agency, a model estimating the likelihood that a discretionary search of a stopped individual recovers contraband.* For the purpose of this analysis, we assume that contraband recovery is the sole motivation for conducting a discretionary search. After subsetting to individuals who were searched for a discretionary reason, such as evidence of a crime or the smell of contraband,† we fit a random forest model predicting contraband recovery based on all recorded factors that an officer could reasonably account for in their decision to search, irrespective of legality.‡ These covariates

*The RIPA dataset also includes instances in which contraband was recovered in plain view (i.e., not as part of a search). However, our risk models specifically estimate the likelihood of contraband recovery following an officer's decision to search.

†If an officer does not have discretion to search a stopped individual, officer-perceived risk of carrying contraband may not factor into the decision to search. So, we do not consider non-discretionary searches when fitting the risk models.

‡We fit the random forest model in `R` using the `ranger` package. We use 128 trees and the default

include the traffic violation or suspected criminal offense that prompted the stop and the basis or bases for conducting the search (e.g., "evidence of a crime" and "contraband in plain view"). In our main analysis, we also include gender and race under the rationale that the stop decision may be affected by additional, risk-relevant and legally permissible factors that gender and race serve as a proxy for, such as socioeconomics. But because this inclusion may raise concerns for disparate treatment violations, in Figure A12, we include an analysis in which we estimate risk without the use of gender and race.* The results are substantively similar.

For each stopped individual, we use the fitted risk model to estimate the probability of recovering contraband from a search—regardless of whether a search was actually conducted.† Of course, contraband recovery from searches is only observed among individuals who were actually searched, so it is impossible to verify the accuracy of the risk model among individuals who were not searched. If there exists an unobserved variable that is correlated with both the search decision and the likelihood of carrying contraband, then our risk estimates will suffer from omitted variable bias [Angrist and Pischke, 2008]. For the purposes of illustration, we proceed under the assumption of no omitted variable bias. In other words, we assume that the decision to search is ignorable: conditional on observed covariates, the search decision is independent of carrying contraband. As a robustness check, Figures A13, A14, and A15 show the results of a sensitivity analysis as proposed in Jung et al. [2023]. For several agencies, we find that estimates of disparate impact are robust to a degree of omitted variable bias comparable to blinding the risk models to the motivating offense of each stop.

Figure 3 shows, for stopped individuals in each agency, the observed probability of being searched, conditional on the estimated risk of carrying contraband. For the majority of the largest agencies, Black and Hispanic individuals were substantially more likely to be searched than white individuals of similar estimated risk. A risk-adjusted regression fit to each agency's data confirms the visual pattern in Figure 3: conditional on estimated risk, both Black and Hispanic individuals were significantly more likely to be searched than white individuals in the majority of agencies (Figure 4). The results of the risk-adjusted regression suggest that the observed adverse impact of searches (see Table A1) is not fully explained by the estimated risk of recovering contraband, which we assume is the primary justification for conducting a search.

For completeness, in the Appendix, we contrast the results from our risk-adjusted regression to measures of disparate impact as obtained through the estimator suggested in

parameters. One could alternatively fit a more complex risk model, such as a neural network, or a less complex model, such as a regularized logistic regression. Table A2 shows all covariates included in the risk models. Figure A8 displays the calibration of each risk model. Figure A9 provides the estimated out-of-sample AUC of the random forest risk model fit to each agency's data.

*For similar reasons, we do not include gender or race when presenting feasible policy alternatives.

†Indicators of criminal activity (e.g., "evidence of a crime") are typically only recorded in the RIPA data when a search was conducted (as a "search basis"). To impute risk for stops in which no search was conducted, we assume none of these factors were present (e.g., if an officer did not conduct a search, we assume they did not see "evidence of a crime"). We make one exception to this assumption: for stops in which contraband was recovered in plain view, without a search, we manually mark the "contraband in plain view" search basis as true. We examine the sensitivity of our results to this inference strategy in the appendix. Figure A17 shows, by agency, the distribution of estimated risk for stopped Black, Hispanic, and white individuals.

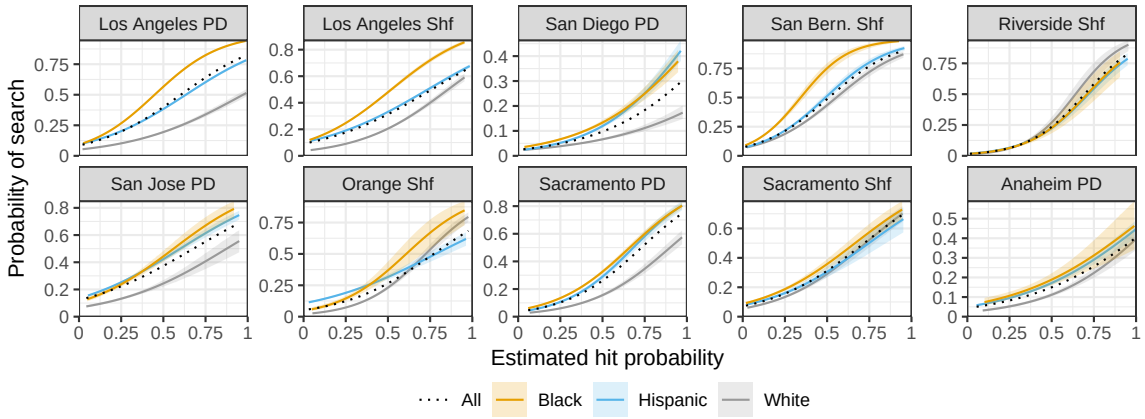

Figure 3: *For individuals stopped by the 10 largest agencies, the probability of being searched as a function of the estimated probability of carrying contraband (i.e., a "hit"), with 95% confidence bands. Hit probability is estimated via a random forest model. In the majority of agencies, Black and Hispanic individuals are substantially more likely to be searched than white individuals with similar estimated risk. Figure A10 expands this figure to include the 50 largest agencies.*

Arnold et al. [2021, 2022]. We observe that estimates are broadly consistent across agencies (Figure A16). We hypothesize that this is because risk distributions across groups are observably similar in most jurisdictions, thus removing concerns arising from inframarginality (Figure A17). There are notable exceptions, however. For instance, whereas risk-adjusted regression estimates suggest unjustified disparities in stops of Hispanic drivers of the Fairfield Police Department, the estimator proposed by Arnold et al. [2021, 2022] does not show similar disparities. Consistent with this divergence, we find that risk distributions in that jurisdiction differ markedly across groups.

4.2 Identifying alternative policies

The coefficients of the risk-adjusted regression models suggest that the adverse impact imposed by search decisions is not *justified* by the estimated risk of carrying contraband. In a last step, we turn to the question of whether there exist implementable alternative search policies that have lower adverse impact and are at least as efficient as the status quo search policy. Although there is not an agreed-upon definition of efficiency with respect to search decisions, for the purpose of this analysis we define a search policy as efficient if it results in the same number of contraband recoveries, in expectation, as the status quo policy without increasing the total number of searches. To guarantee officers are not required to do additional work under any of our proposals, we restrict ourselves to policies that do not increase the total number of searches. It is possible, however, that police agencies and courts might deem it acceptable to increase the space of policies to consider in order to find one that imposes less adverse impact.

To identify an initial efficient threshold policy, we sort all n stopped individuals in descending order by their estimated risk. We then iterate through possible risk thresholds,

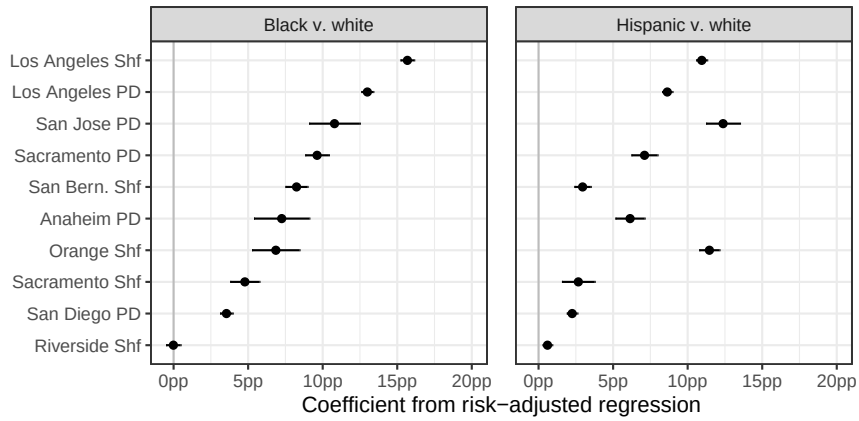

Figure 4: *Coefficients on the race and ethnicity terms from risk-adjusted regression models fit separately to each of the 10 largest agencies, with 95% confidence intervals. In the majority of agencies, both stopped Black and stopped Hispanic individuals are significantly more likely to be searched than stopped white individuals of similar estimated risk, suggesting that the observed adverse impact of searches (see Table A1) is not explained by the risk of carrying contraband. Figure A11 shows the same coefficients for the 50 largest agencies.*

where the k individuals above each risk threshold t are assumed to be searched, and the $n-k$ remaining individuals are not. We measure the adverse impact of this policy as the ratio of the resulting per capita search rates for Black and white individuals. To sweep out the remaining threshold policies, we iterate over lower values of t until k is approximately the same as the total number of individuals searched by the status quo policy. At each iteration, we calculate the adverse impact and the expected number of contraband recoveries.*

One might, however, argue that these threshold decision rules—which involve complex risk estimation using a random forest model—are not practically implementable. To address this concern, we follow Goel et al. [2016] and construct more readily implementable decision rules (See "Constructing the simple rule" in the Appendix for the rule construction process). These simple rules account only for the traffic violation or suspected offense that motivated the stop, the city in which the stop occurred, whether contraband is in plain view, and whether there is evidence of a crime. To use these simple rules, officers would only need to add up two small integers, and compare the result to a threshold unique to each combination of city and traffic violation or suspected offense.

Disparate impact law stipulates that the benchmark against which one should measure decision rates consists of those affected by the decision, or those who could be affected by a change in the way the decision is determined [Carpenter v. Boeing Co., 456 F.3d 1183 (10th Cir. 2006); Hous. Invs., Inc. v. City of Clanton, Ala., 68 F. Supp. 2d 1287 (M.D. Ala. 1999)]. Following these guidelines, we calculate adverse impact using two reasonable benchmark populations. First, we calculate the ratio of per capita search rates, which are computed by dividing the number of searches for a given race or ethnicity by the entire population of that group within the jurisdiction patrolled by the agency.† This

*We calculate this expectation by summing the estimated risk of the k searched individuals.

†City-level demographics are sourced from the 2022 Population and Housing Unit Estimates of the U.S.

population-level benchmark is intended to be representative of all individuals who could have been searched by law enforcement agencies. Second, we also compute adverse impact as the ratio of stop-level search rates, which are computed by dividing the total number of searched individuals in each group by the total number of stopped individuals in each group. The second benchmark follows from a narrower perspective of disparate impact in search decisions where the affected group consists of just those who were stopped.* The main results use the ratio of per capita search rates, with results based on the ratio of stop-level search rates in the appendix (Figure A21).

Figure 5 shows the adverse impact resulting from the threshold policies derived from the iterative process outlined above. The dotted line in each panel represents policies where search decisions are determined by the simple rule.† The dotted line begins at the threshold where the expected number of contraband recoveries is the same as the actual number of contraband recoveries. This is the policy with the highest threshold that is arguably as efficient as the status quo policy, so we do not show policies with higher thresholds (i.e., fewer searches). Analogously, we do not show policies with lower thresholds than the policy that searches the same number of individuals as the status quo, which corresponds to the far right end of each dotted line. As we sweep across smaller thresholds, the total number of allowed searches increases.

The x-axis shows, for each policy, the number of searches conducted (k) divided by the total number of searches observed in the real data (n). Finally, for comparison, the solid arrow on the right side of each panel indicates the adverse impact observed under the status quo search policy, measured as the ratio of per capita search rates among each minority group and white individuals (see Table A1).‡ For the majority of agencies, there exists a simple rule threshold policy with lower adverse impact on stopped Black individuals and Hispanic individuals. Further, all of these policies are able to recover at least as many weapons as the status quo, in expectation, with fewer searches. These results show the existence of an equally efficient policy with lower adverse impact, arguably meeting the plaintiff's burden under the third step of a disparate impact claim.

5 Discussion

In practice, risk-based approaches to disparate impact are only applicable in certain settings. First, there must be a measurable indicator of a successful decision. As an example, consider the pretrial setting. In most jurisdictions, the primary justification for pretrial detention is minimizing the risk of failing to appear or committing new criminal activity. The existence of a pretrial violation is a concrete way to assess whether a release decision is "successful". In other domains, such as college admissions, it is not immediately clear how to denote a successful decision. Second, one must be able to estimate risk accurately. This typically means that decision rates must be high enough such that there exists a sufficient number of

Census Bureau. County-level demographics are sourced from the 2020 U.S. Census.

*Ultimately, the scope of a hypothetical disparate impact claim would inform the appropriate choice of reference group.

†Figure A20 shows the same results using a random forest risk model instead of the simple rule.

‡Figure A21 shows the same results using the ratio of stop-level search rates as the measure of adverse impact.

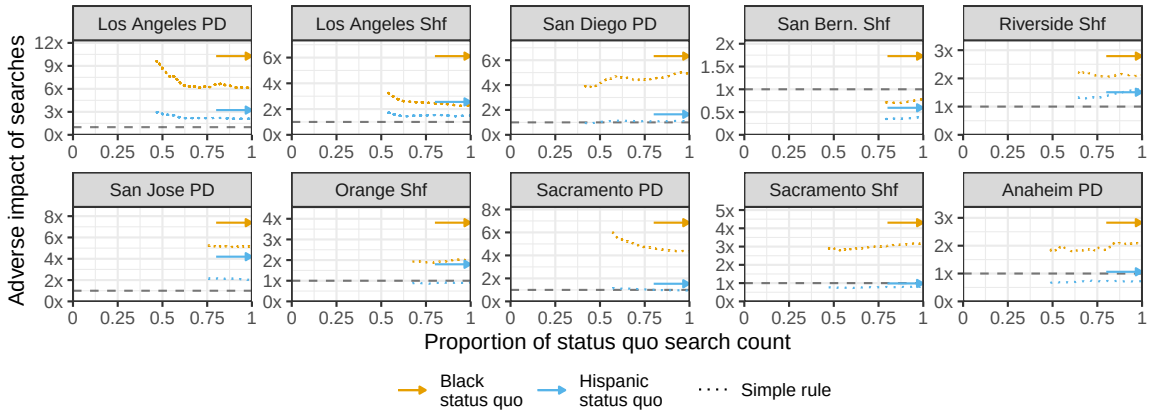

Figure 5: *For the 10 largest agencies, estimated adverse impact under hypothetical threshold policies that, compared to the status quo, recover at least as much contraband and result in no additional searches. Adverse impact is measured as the ratio of per capita search rates. Contraband recovery risk is estimated via a simple rule. For the majority of agencies, there exist threshold policies (dotted lines) with lower adverse impact on stopped Black individuals and Hispanic individuals. Figure A19 show the same results for the 50 largest agencies.*

individuals from which to estimate risk. Additionally, as accurate risk estimation often rests on the strength of the ignorability assumption, the fitted risk model should incorporate as many of the variables observed by the decision maker as possible. If there are unobservable variables that are highly predictive of both the decision itself and the success of the decision, the risk model may suffer from severe omitted variable bias. Finally, the proposed risk-based alternative policies must be implementable. For example, the decision to search a pedestrian may be made in a matter of seconds, so even a simple rule could be deemed as impossible to realistically implement. For less time-constrained decisions, such as pretrial detention or tax auditing, risk estimates can be generated well in advance of decisions.

In addition, we note that, in this study, we take disparate impact law as given* and consider empirical strategies in relation to current legal analysis. But we believe the current law on disparate impact has many shortcomings itself. Among others, disparate impact law's focus on raw disparities can lead decision makers to forego policies that are ultimately favorable to the minority group. For instance, a policy that is strictly beneficial to both the minority and the majority group, but that benefits the majority group more than the minority group, would not need to be enacted under disparate impact law because it *increases* the disparities between the groups.

To illustrate with a numeric example, consider a hypothetical scenario under which the current policy has police officers search pedestrians if they made 'furtive movements.' Under this policy, the officer stops 100 of 10,000 Black citizens a year, and 1,000 of 100,000 white citizens. An analysis shows that, although officers do not act with discriminatory intent, 'furtive movements' is not predictive of weapon recovery. Removing this requirement would thus reduce the number of Black citizens stopped by 50, and the number of white citizens

*Noting again that the broader applicability of disparate impact law to policing is speculative.

stopped by 600, without meaningfully affecting the weapon recovery rate. In this scenario, the current policy has a search rate of 1% for both Black and white citizens. Under the new policy, the search rate is reduced to 0.5% for Black citizens and 0.4% for white citizens. But although the new policy decreases the absolute number of both Black and white citizens who are searched, it *increases* the relative disparity between Black and white citizens from 0.0 to 0.1 percentage points. Under disparate impact law, the new policy need not be implemented, given that it does not decrease the disparities between the two groups.

The example helps clarify the focus of disparate impact law, and how it might differ from other welfare perspectives on fairness. Disparate impact law is primarily concerned with unjustified, differential treatment between the majority and the minority group. However, it is not a mandate to improve the welfare of the minority group, even if that can be done in a costless way. From a welfarist perspective, this might seem problematic, especially in settings where there is no budget constraint.

Additionally, the reference population from which action rates are calculated should, in theory, consist of those who are subjected to the practice in question [Carpenter v. Boeing Co., 456 F.3d 1183 (10th Cir. 2006); Hous. Invs., Inc. v. City of Clanton, Ala., 68 F. Supp. 2d 1287 (M.D. Ala. 1999)]. In practice, though, it is often unclear what the relevant reference population should be. Furthermore, data for certain reference populations may be inaccessible. For example, in the case of lending, one might propose a reference population of all eligible individuals who applied for a loan from the institution in question. However, it appears to us that a more suitable population would be all individuals who *would have* been eligible for a loan, regardless of whether they actually applied. But, the size of this larger group may not be estimable, in which case the smaller group would be an appropriate reference population so long as it is sufficiently representative of the affected individuals [Frazier v. Consol. Rail Corp., 851 F.2d 1447 (D.C. Cir. 1988)]. Courts have also permitted reference populations that subsume the affected population, once again so long as the larger population is sufficiently representative [E.E.O.C. v. Joint Apprenticeship Comm. of Joint Indus. Bd. of Elec. Indus., 186 F.3d 110 (2d Cir. 1999)].

6 Conclusion

In this paper, we have discussed statistical approaches for assessing disparate impact. Our analysis suggests that recent estimators centered on error rates capture neither legal nor normative notions of disparate impact. While risk-adjusted regression can help document the existence of unjustified disparities, a concrete, optimal alternative policy can be derived by sorting individuals based on their estimated risk and defining a decision threshold. As we have shown for the example of search decisions by California law enforcement agencies in 2022, this approach relies on analysts to formulate alternative, less disparate, implementable policies even in scenarios where decision makers have constrained information or time. We hope that this research will positively contribute towards a current trend to broaden conceptions of discrimination in empirical research.

References

Joshua D Angrist and Jörn-Steffen Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2008.

David Arnold, Will Dobbie, and Peter Hull. Measuring racial discrimination in algorithms. In *AEA Papers and Proceedings*, volume 111, pages 49–54, 2021.

David Arnold, Will Dobbie, and Peter Hull. Measuring racial discrimination in bail decisions. *American Economic Review*, 112(9):2992–3038, 2022.

Ronen Avraham, Kyle D Logue, and Daniel Schwarcz. Understanding insurance antidiscrimination law. *Southern California Law Review*, 87:195, 2013.

Ian Ayres. Outcome tests of racial disparities in police practices. *Justice Research and Policy*, 4(1-2):131–142, 2002.

Ian Ayres. Three tests for measuring unjustified disparate impacts in organ transplantation: The problem of "included variable" bias. *Perspectives in Biology and Medicine*, 48(1): 68–S87, 2005.

Ian Ayres. Testing for discrimination and the problem of "included variable" bias. *Working paper*, 2010.

E Jason Baron, Joseph J Doyle Jr, Natalia Emanuel, Peter Hull, and Joseph P Ryan. Racial Discrimination in Child Protection. 2023.

Robert Bartlett, Adair Morse, Nancy Wallace, and Richard Stanton. Algorithmic discrimination and input accountability under the civil rights acts. *Berkeley Tech. LJ*, 36:675, 2021.

Gary S Becker. *The Economics of Discrimination*. 1957.

Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44, 2021.

J Aislinn Bohren, Peter Hull, and Alex Imas. Systemic Discrimination: Theory and Measurement. 2022.

Stephanie Bornstein. Reckless Discrimination. *California Law Review*, pages 1055–1110, 2017.

George S Bridges and Sara Steen. Racial disparities in official assessments of juvenile offenders: Attributional stereotypes as mediating mechanisms. *American Sociological Review*, pages 554–570, 1998.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91, 2018.

William Cai, Johann Gaebler, Justin Kaashoek, Lisa Pinals, Samuel Madden, and Sharad Goel. Measuring racial and ethnic disparities in traffic enforcement with large-scale telematics data. *PNAS Nexus*, 1(4), 2022.

Dallas Card and Noah A Smith. On consequentialism and fairness. *Frontiers in Artificial Intelligence*, 3:34, 2020.

Anupam Chander. The racist algorithm. *Michigan Law Review*, 115:1023, 2016.

Alex Chohlas-Wood, Madison Coots, Sharad Goel, and Julian Nyarko. Designing equitable algorithms. *Nature Computational Science*, 3(7):601–610, 2023.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163, 2017.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806, 2017.

Sam Corbett-Davies, J Gaebler, Hamed Nilforoshan, Ravi Shroff, and Sharad Goel. The measure and mismeasure of fairness. *Journal of Machine Learning Research*, 2023.

Stephen Demuth. Racial and ethnic differences in pretrial release decisions and outcomes: A comparison of hispanic, black, and white felony arrestees. *Criminology*, 41(3):873–908, 2003.

Stephanie Holmes Didwania. Discretion and disparity in federal detention. *Northwest. Univ. Law Rev.*, 115:1261, 2020.

Edward C Dillon Jr, Juan E Gilbert, Jerlando FL Jackson, and LJ Charleston. The state of African Americans in computer science—the need to increase representation. *Computing Research News*, 21(8):2–6, 2015.

Ellen A Donnelly and John M MacDonald. The downstream effects of bail and pretrial detention on racial disparities in incarceration. *J. Crim. l. & CriminoloGY*, 108:775, 2018.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226, 2012.

Hadi Elzayn, Evelyn Smith, Thomas Hertz, Arun Ramesh, Jacob Goldin, Daniel E Ho, and Robin Fisher. Measuring and mitigating racial disparities in tax audits. 2023.

Johann Gaebler, William Cai, Guillaume Basse, Ravi Shroff, Sharad Goel, and Jennifer Hill. A causal framework for observational studies of discrimination. *Statistics and Public Policy*, 9(1):26–48, 2022.

Sharad Goel, Justin M Rao, and Ravi Shroff. Precinct or prejudice? Understanding racial disparities in New York City's stop-and-frisk policy. *Annals of Applied Statistics*, 10(1): 365–394, 2016.

Steven N Goodman, Sharad Goel, and Mark R Cullen. Machine learning, health disparities, and causal reasoning. *Annals of Internal Medicine*, 169(12):883–884, 2018.

Joshua Grossman, Sabina Tomkins, Lindsay C Page, and Sharad Goel. The Disparate Impacts of College Admissions Policies on Asian American Applicants. *Scientific Reports*, 2024.

Ben Grunwald, Julian Nyarko, and John Rappaport. Police agencies on Facebook overreport on Black suspects. *Proceedings of the National Academy of Sciences*, 119(45), 2022.

Cheryl I Harris. Limiting Equality: The Divergence and Convergence of Title VII and Equal Protection. *U. Chi. Legal F.*, page 95, 2014.

Paul Heaton, Sandra Mayson, and Megan Stevenson. The Downstream Consequences of Misdemeanor Pretrial Detention. *Stan. L. Rev.*, 69:711, 2017.

Brian Hedden. On Statistical Criteria of Algorithmic Fairness. 2021.

Deborah Hellman. Measuring algorithmic fairness. *Virginia Law Review*, 106(4):811–866, 2020.

Mitchell Hoffman, Lisa B Kahn, and Danielle Li. Discretion in hiring. *The Quarterly Journal of Economics*, 133(2):765–800, 2018.

Lily Hu and Issa Kohler-Hausmann. What's sex got to do with machine learning? In *Proceedings of the 2020 ACM Conference on Fairness, Accountability, and Transparency*, 2020.

Aziz Huq. Racial equity in algorithmic criminal justice. *Duke Law Journal*, 68, 2019.

Aziz Z Huq. What is Discriminatory Intent? *Cornell L. Rev.*, 103:1211, 2017.

Kosuke Imai, Zhichao Jiang, D James Greiner, Ryan Halen, and Sooahn Shin. Experimental evaluation of algorithm-assisted human decision-making: application to pretrial public safety assessment. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 186(2):167–189, 02 2023.

Jongbin Jung, Sam Corbett-Davies, Johann Gaebler, Ravi Shroff, and Sharad Goel. Mitigating included- and omitted-variable bias in estimates of disparate impact. *arXiv preprint arXiv:1809.05651*, 2023.

Maximilian Kasy and Rediet Abebe. Fairness, equality, and power in algorithmic decision-making. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 576–586, New York, NY, USA, 2021. Association for Computing Machinery.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *Proceedings of Innovations in Theoretical Computer Science (ITCS)*, 2017.

Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Cass R Sunstein. Discrimination in the age of algorithms. *Journal of Legal Analysis*, 10:113–174, 2018.

Sandra G Mayson. Bias in, bias out. *The Yale Law Journal*, 128(8):2218–2300, 2019.

Melissa D McCradden, Shalmali Joshi, Mjaye Mazwi, and James A Anderson. Ethical limitations of algorithmic fairness solutions in health care machine learning. *The Lancet Digital Health*, 2(5):e221–e223, 2020.

Christi Metcalfe and Ted Chiricos. Race, plea, and charge reduction: An assessment of racial disparities in the plea process. *Justice Quarterly*, 35(2):223–253, 2018.

David Benjamin Oppenheimer. Negligent Discrimination. *U. Pa. L. Rev.*, 141:899, 1992.

Jessica K Paulus and David M Kent. Predictably unequal: understanding and addressing concerns that algorithmic clinical prediction may increase health disparities. *NPJ Digital Medicine*, 3(1):1–8, 2020.

M Marit Rehavi and Sonja B Starr. Racial disparity in federal criminal sentences. *Journal of Political Economy*, 122(6):1320–1354, 2014.

Maya Sen and Omar Wasow. Race as a bundle of sticks: Designs that estimate effects of seemingly immutable characteristics. *Annual Review of Political Science*, 19(1):499–522, 2016.

Camelia Simoiu, Sam Corbett-Davies, and Sharad Goel. The problem of infra-marginality in outcome tests for discrimination. *The Annals of Applied Statistics*, 11(3):1193–1216, 2017.

Alisa Tiwari. Disparate-impact liability for policing. *The Yale Law Journal*, pages 252–306, 2019.

John Wooldredge. Distinguishing race effects on pre-trial release and sentencing decisions. *Justice Quarterly*, 29(1):41–75, 2012.

Michael Zanger-Tishler, Julian Nyarko, and Sharad Goel. Risk scores, label bias, and everything but the kitchen sink. *Science Advances*, 2024.

A Appendix

Table of Contents

Figure A1 is an extension of Figure 2 to a setting where risk is continuously distributed. Specifically, risk is parameterized by a beta distribution with a fixed variance and mean between 0 and 1.

Figures A2, A3, and A4 are excerpts from the RIPA data collection form.

Figure A5 shows search and arrest rates for each stop reason.

Table A1 shows adverse impact across the 50 largest agencies in the RIPA data.

Table A2 lists the covariates included in each agency's risk model.

Figure A6 shows why seven of the largest agencies are excluded from the analysis.

Figure A7 shows the distribution of variable importance for the covariates included in each agency's random forest risk model.

Figure A8 shows calibration plots for each agency's risk model.

Figure A9 shows the out-of-sample AUC of each agency's risk model.

Figure A10 shows search rates as a function of estimated risk for each agency, analogously to Figure 3.

Figure A11 shows the race and ethnicity coefficients from the risk-adjusted regression models fit to each agency's data, analogously to Figure 4.

Figure A12 shows robustness checks for the risk-adjusted regression models fit to each agency's data.

Figures A13, A14, and A15 show the results of sensitivity analyses of the risk-adjusted regression coefficients for each agency.

Figures A16 and A18 compare the risk-adjusted regression coefficients to the estimates derived from the Arnold et al. [2021] measure of disparate impact.

Table A17 shows the distribution of estimated risk for stopped individuals.

"Constructing the simple rule" outlines the process used to construct the simple rule risk models used in Figure 5.

Figures A19, A20, and A21 show the adverse impact observed under hypothetical threshold policies, analogously to Figure 5.

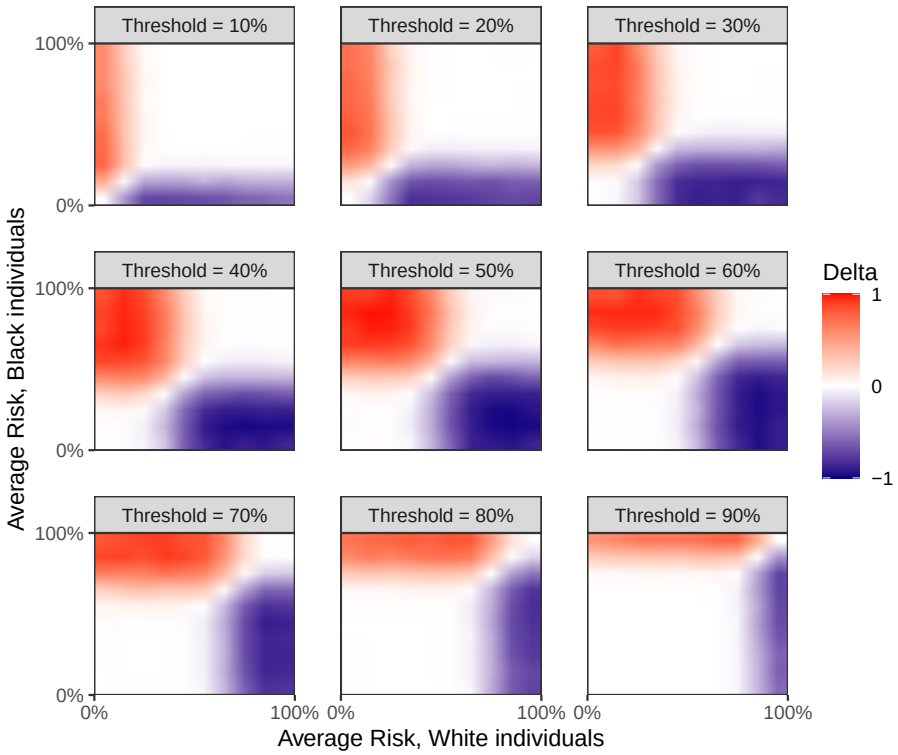

Figure A1: *Extension of the scenario in Figure 1 to continuous risk. In this example, risk follows a beta distribution with a mean between 0 and 1 and a fixed variance of 0.01. Each panel shows values of* Δ *(Delta) calculated under a fixed search threshold for synthetic groups of Black and white stopped individuals, with risk values randomly generated from a beta distribution with the mean specified on the corresponding axis. As in Figure 1, each panel has a white diagonal line, indicating that the* Δ *measure is correctly 0 when risk distributions are identical. When the mean of both risk distributions is far from the search threshold,* Δ *is also close to 0. However, when the mean of either risk distribution is close to the search threshold, the calculated value of* Δ *is more sensitive to small changes in the shape of the risk distribution(s) whose mean is close to the search threshold.*

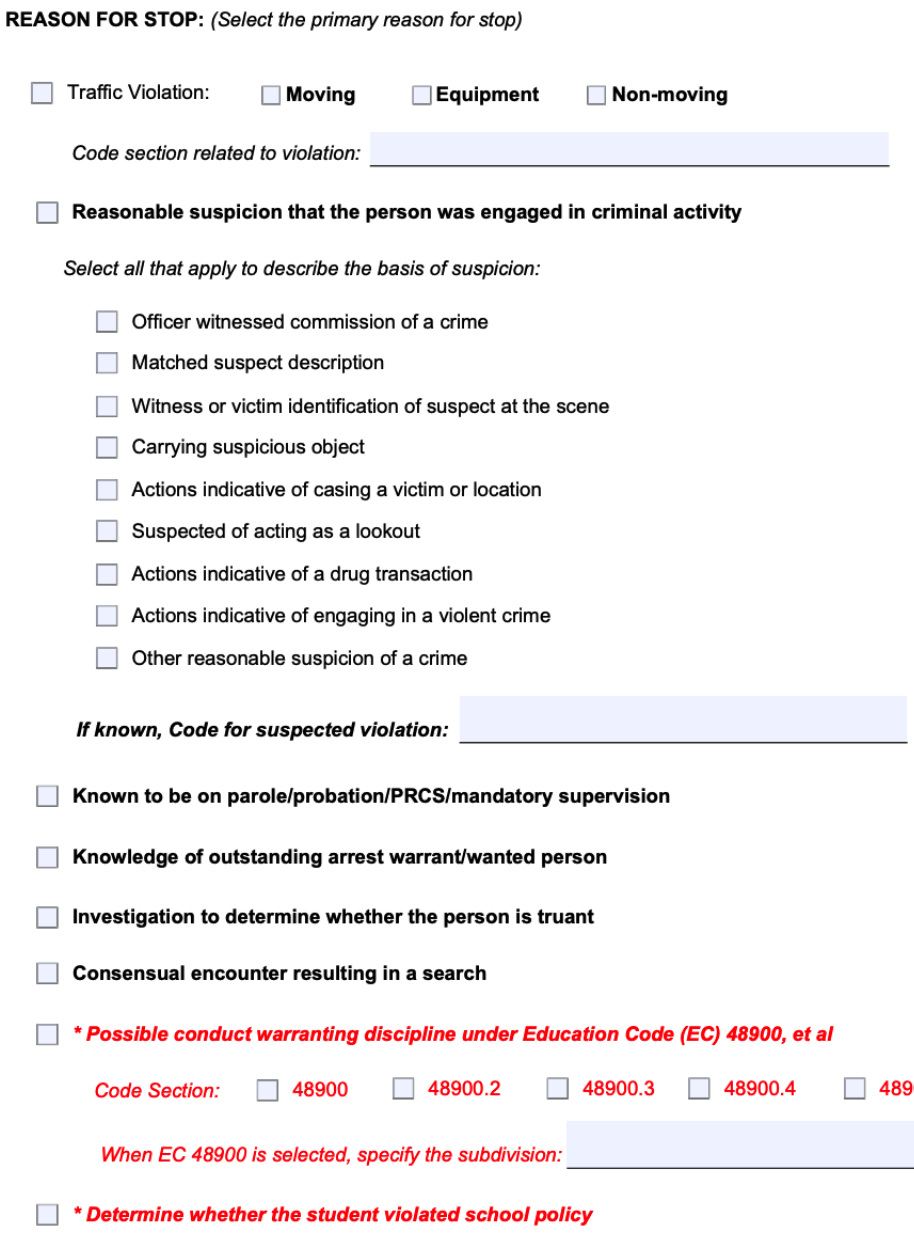

Figure A2: *Stop reasons listed on the RIPA data collection form. Text position is altered slightly for readability.*

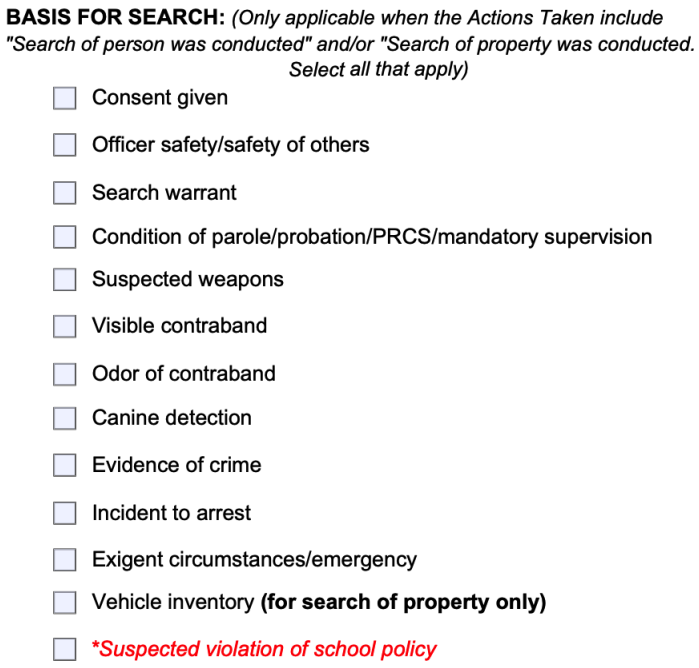

Figure A3: *Search bases listed on the RIPA data collection form. Text position is altered slightly for readability.*

CONTRABAND/EVIDENCE DISCOVERED (IF ANY): *(Include any items discovered in plain view or as the result of a search)*
Select all that apply:

☐ None	☐ Drugs/narcotics	☐ Suspected stolen property
☐ Firearm(s)	☐ Alcohol	☐ Cell phone(s) or electronic devices(s)
☐ Ammunition	☐ Money	☐ Other contraband or evidence
☐ Weapon(s) other than firearm	☐ Drug Paraphernalia	

Figure A4: *Types of contraband listed on the RIPA data collection form. Text position is altered slightly for readability.*

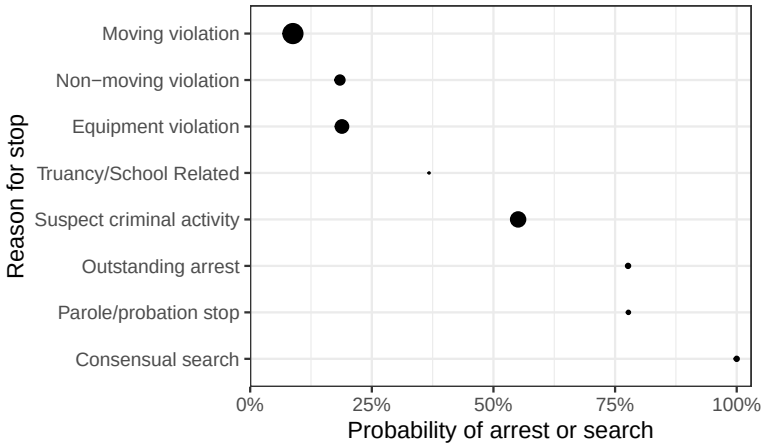

Figure A5: *For all stops recorded by the 50 law enforcement agencies, the proportion of stops resulting in a search or arrest, disaggregated by stop reason. The size of the points corresponds to the number of stops prompted by each reason. Stops initiated due to an outstanding arrest, parole, or probation are excluded from the main analysis, as are consensual searches and stops initiated due to school policy.*

Agency	Stops	B (pop)	H (pop)	B (stop)	H (stop)
Los Angeles PD	320,702	11.8	3.8	2.7	2.1
Los Angeles Shf	273,000	6.5	2.7	2.1	1.8
San Diego PD	92,898	6.8	1.8	1.8	1.4
San Bern. Shf	81,548	1.7	0.6	1.1	0.9
Riverside Shf	68,946	2.8	1.5	1.2	1.2
San Jose PD	39,776	7.4	4.3	1.9	1.8
Orange Shf	38,916	4.2	1.8	2.0	2.2
Sacramento PD	38,113	7.4	1.7	2.1	1.7
Sacramento Shf	24,902	4.2	1.0	1.3	1.2
Anaheim PD	21,349	2.7	1.1	1.4	1.3
Irvine PD	20,335	4.3	3.3	1.3	1.1
Ventura PD	20,121	4.1	1.5	1.1	1.2
Burbank PD	19,930	7.7	3.3	1.8	1.7
San Diego Shf	18,482	1.9	1.4	1.7	1.9
Fontana PD	18,255	1.5	0.9	0.9	0.9
Santa Rosa PD	18,189	6.7	2.1	1.3	1.3
Escondido PD	16,111	4.4	1.1	1.3	1.1
Riverside PD	16,090	4.7	1.1	2.0	1.2
San Mateo PD	15,136	7.6	3.3	2.2	1.7
Placer Shf	14,853	8.8	1.6	1.4	1.2
Redding PD	14,789	2.6	0.5	0.9	0.9
Oxnard PD	14,360	3.1	1.4	1.5	1.5
San Mateo Shf	14,315	5.1	2.0	2.5	1.6
San Fran. PD	14,273	8.8	2.2	1.7	1.3
Oceanside PD	14,255	2.7	1.7	1.1	1.5
Oakland PD	13,941	6.9	3.0	1.2	1.0
Rialto PD	13,421	1.5	0.6	1.3	0.9
Costa Mesa PD	13,103	5.6	1.0	1.6	1.1
Coronado PD	12,825	9.3	6.7	1.7	2.3
Chino PD	12,575	1.0	1.2	0.7	0.9
Fresno Shf	12,140	2.5	1.3	0.9	1.0
Pasadena PD	11,968	9.2	2.5	3.6	2.2
Alameda Shf	11,595	4.7	2.1	1.7	1.2
Clovis PD	11,336	3.8	1.1	1.1	1.1
Santa Ana PD	11,078	3.3	1.0	1.3	0.9
Hanford PD	10,680	3.1	1.8	1.4	1.1
Orange PD	10,244	5.2	2.1	1.6	1.5
BART PD	10,032	11.2	1.3	1.1	1.1
Folsom PD	10,011	3.8	1.2	2.3	1.5
Livermore PD	9,730	11.9	2.5	1.4	1.2
Htg. Bch. PD	9,634	4.6	1.4	1.1	1.1
Vacaville PD	9,428	6.6	1.5	1.8	1.3
Visalia PD	9,424	5.0	1.9	2.0	1.3
Roseville PD	9,274	6.8	1.1	1.2	1.1
San Joaquin Shf	9,218	1.6	0.9	1.0	1.1
Petaluma PD	9,213	7.5	2.4	1.7	1.3
Fairfield PD	9,105	3.1	1.0	1.3	1.0
Glendale PD	8,985	9.9	3.5	2.2	2.1
Carlsbad PD	8,946	6.9	2.2	1.2	1.3
Pacifica PD	8,766	4.6	1.9	1.4	1.2

Table A1: *For the 50 agencies with the most stops recorded in 2022, adverse impact of search decisions for Black (B) and Hispanic (H) individuals, relative to white individuals. Adverse impact is defined at the population-level ("pop") by the ratio of race-specific, per capita search rates. Per capita search rates are computed as the number searched over the population of the jurisdiction served by the agency. For example, a value of 2 for "B (pop)" implies that the per capita search rate for Black individuals was twice as high as that of white individuals. Adverse impact is defined at the stop-level ("stop") by the ratio of stop-level search rates. Stop-level search rates are computed as the number searched over the number stopped. CHP: California Highway Patrol; PD: Police Department; Shf: County Sheriff; Bern: Bernardino; Htg Bch: Huntington Beach; BART: Bay Area Rapid Transit.*

Variable	Description	Possible values
Race/ethnicity	Race or ethnicity of the driver, as perceived by the officer. Approximately 99% of individuals in the data are classified into a single race or ethnicity. The remaining 1% are considered 'Hispanic' if they are classified as Hispanic along with any other race or ethnicity, 'Black' if they are classified as Black and any other race or ethnicity other than Hispanic, 'White' if they are classified as 'White' and no other race or ethnicity, and 'Other' otherwise.	White | Black | Hispanic | Other
Gender	Gender of the driver, as perceived by the officer. Approximately 99.7% of individuals in the data are perceived as either male or female. Due to small sample size, the remaining 0.3% of individuals are excluded from the analysis.	Male | Female
Reason for stop	Reason for the stop, as recorded by the officer. Only a single reason can be recorded.	Suspected criminal activity | Equipment violation | Moving violation | Non-moving violation
RAS factors	Reasonable articulable suspicion (RAS) factors recorded by the officer if criminal activity is suspected. Multiple factors may be recorded.	Officer witnessed a crime | Suspect matched description | Witness or victim identification of suspect | Carrying suspicious object | Casing a victim or location | Acting as a lookout | Actions indicative of drug transaction | Actions indicative of violent crime | Other
Search basis	If a search is carried out by the officer, the recorded basis for the search. Multiple bases may be recorded.	Contraband in plain view | Odor of contraband | Officer safety or safety of others | Suspected weapon | Evidence of crime | Emergency | Canine detection of contraband | School policy
Traffic offense by ped.	Whether the traffic offense that led to the stop was likely carried out by a pedestrian (e.g., jaywalking).	Yes | No
RAS offense by driver	Whether the reasonable suspicion offense that led to the stop was likely carried out by a driver. For example, driving under the influence.	Yes | No
Is call for service	Whether the stop was in response to a call for service.	Yes | No
Multi-person encounter	Whether more than one individual was recorded as stopped. The main analysis considers only the first person listed.	Yes | No
Motivating offense FEs	Fixed effects for the motivating offense that led to the stop.	Robbery | Assault | Burglary | Theft | DUI | Speeding | Etc.
City FEs	Fixed effects for the city in which the stop took place.	Los Angeles | San Diego | San Jose | San Francisco | Etc.
Month FEs	Fixed effects for the month of the stop.	0-11
Weekday FEs	Fixed effects for the day of the week on which the stop took place.	0-6
Hour FEs	Fixed effects for the hour of the day in which the stop took place.	0-23

Table A2: *Covariates included in each agency's random forest risk model used to estimate the likelihood of recovering contraband from a discretionary search.*

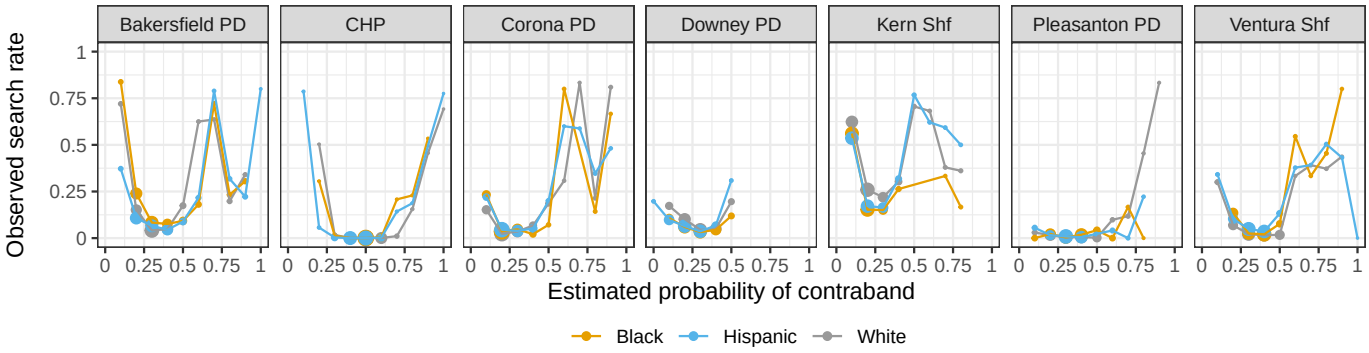

Figure A6: *Large agencies with invalid risk models. These agencies are larger or as large as the 50 agencies included in the expanded analysis. If the estimated risk of recovering contraband is indeed the main motivation for carrying out a discretionary search, then discretionary search rates should increase monotonically as a function of estimated risk. If not, the risk measure may not be an appropriate measure of qualification for a discretionary search. There may be, for example, an agency-specific requirement for conducting a search that is not recorded in the RIPA data. For these agencies, search rates do not increase monotonically with estimated risk, so these agencies are excluded from the main analysis.*

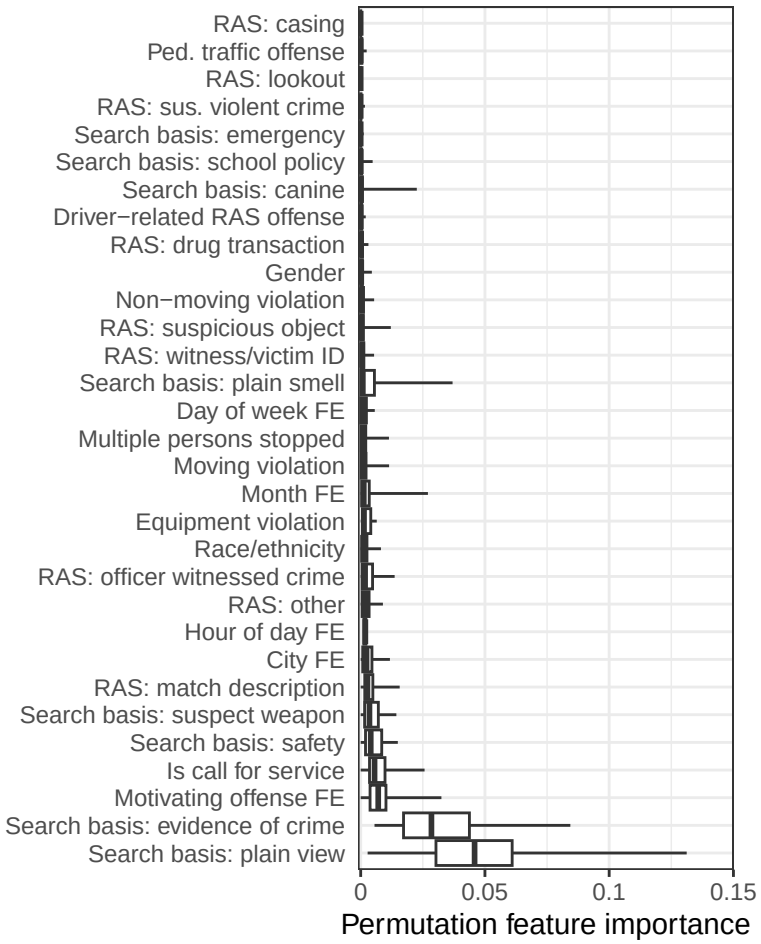

Figure A7: *Boxplots of the permutation feature importance values for each covariate included in the agency-specific random forest risk models. The boxplots show the distribution of feature importance values across all 50 agencies. In most jurisdictions, plain view contraband and evidence of a crime are by far the most influential features for predicting contraband recovery from a discretionary search of a stopped individual.*

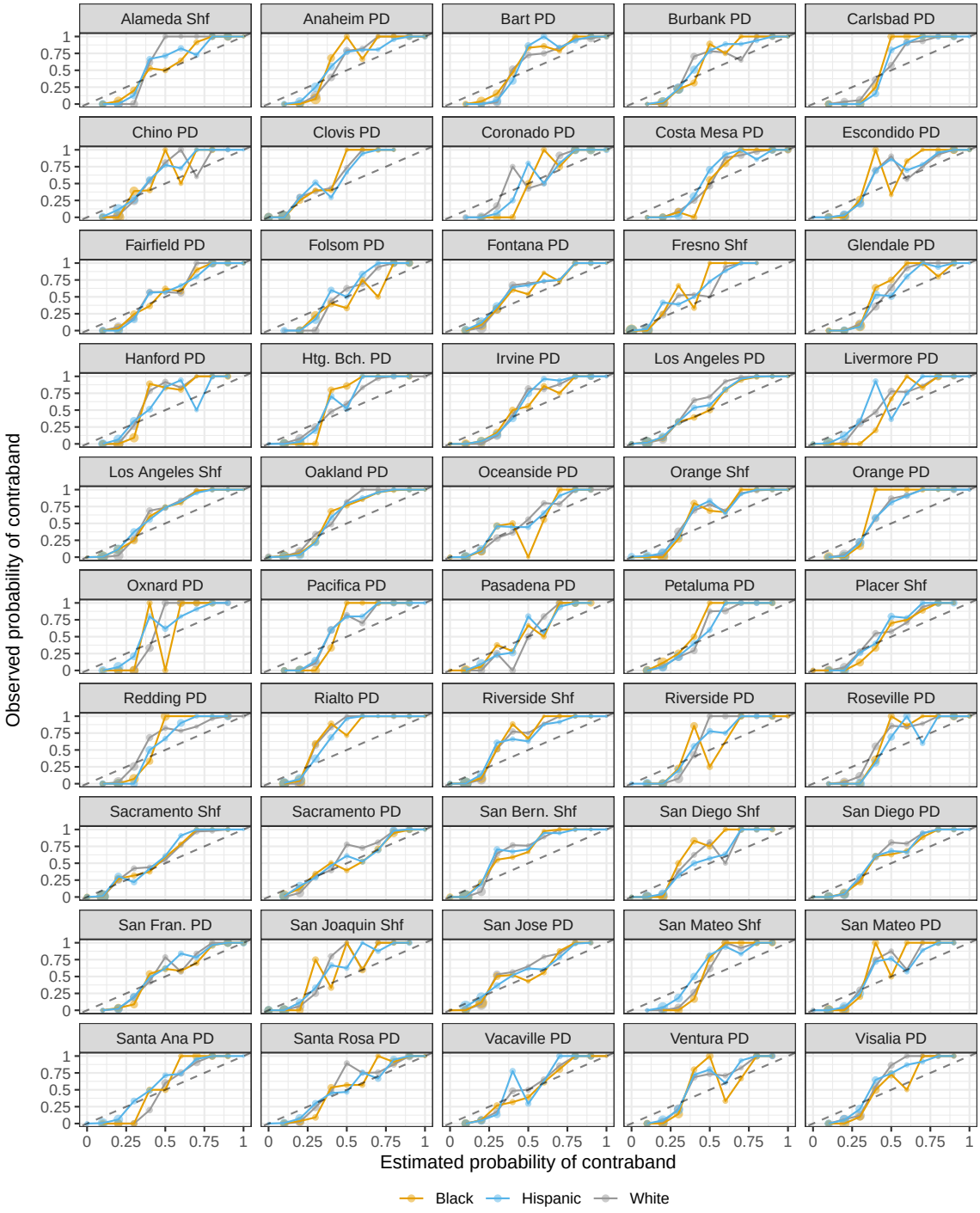

Figure A8: *Calibration plots for the contraband-carrying risk models fit to each of the 50 largest agencies. Larger point sizes correspond to more observations. For a well-fitted risk model, the observed probability of carrying contraband should be monotonic as a function of the estimated probability of carrying contraband. Monotonicity ensures that thresholded search policies will return the same results regardless of whether the model is recalibrated to fall along the parity line. The monotonicity requirement is approximately met by most agencies and race/ethnicity groups.*

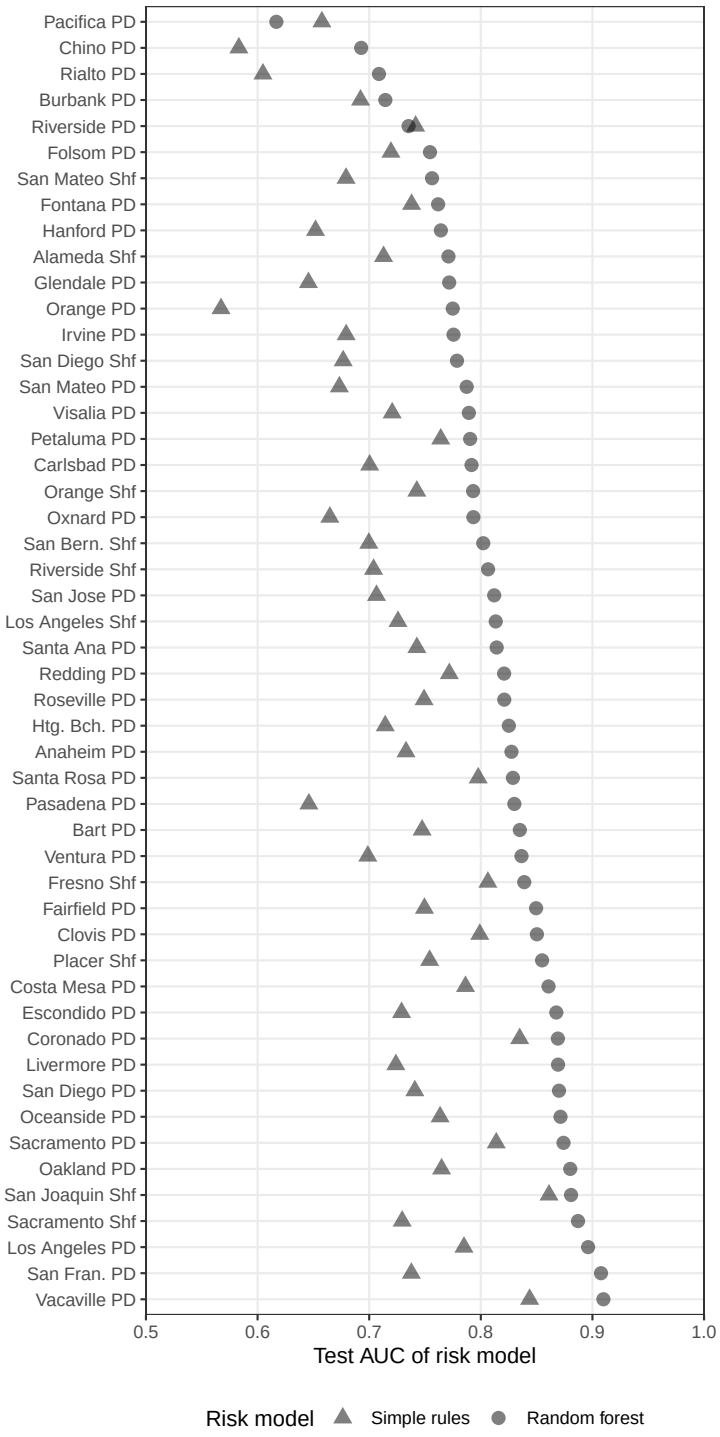

Figure A9: *Estimated out-of-sample AUC for the random forest and simple rules risk models for the 50 largest agencies. AUC is calculated with an 80/20 train/test split. Performance of the random forest risk models is moderate-to-strong across agencies, with almost all agencies having an AUC above 0.7, and the majority having an AUC above 0.8. For most agencies, the simple rules model performs worse than the random forest model, though the performance of the simple rules models exceed 0.65 in almost all jurisdictions, with the majority exceeding 0.7.*

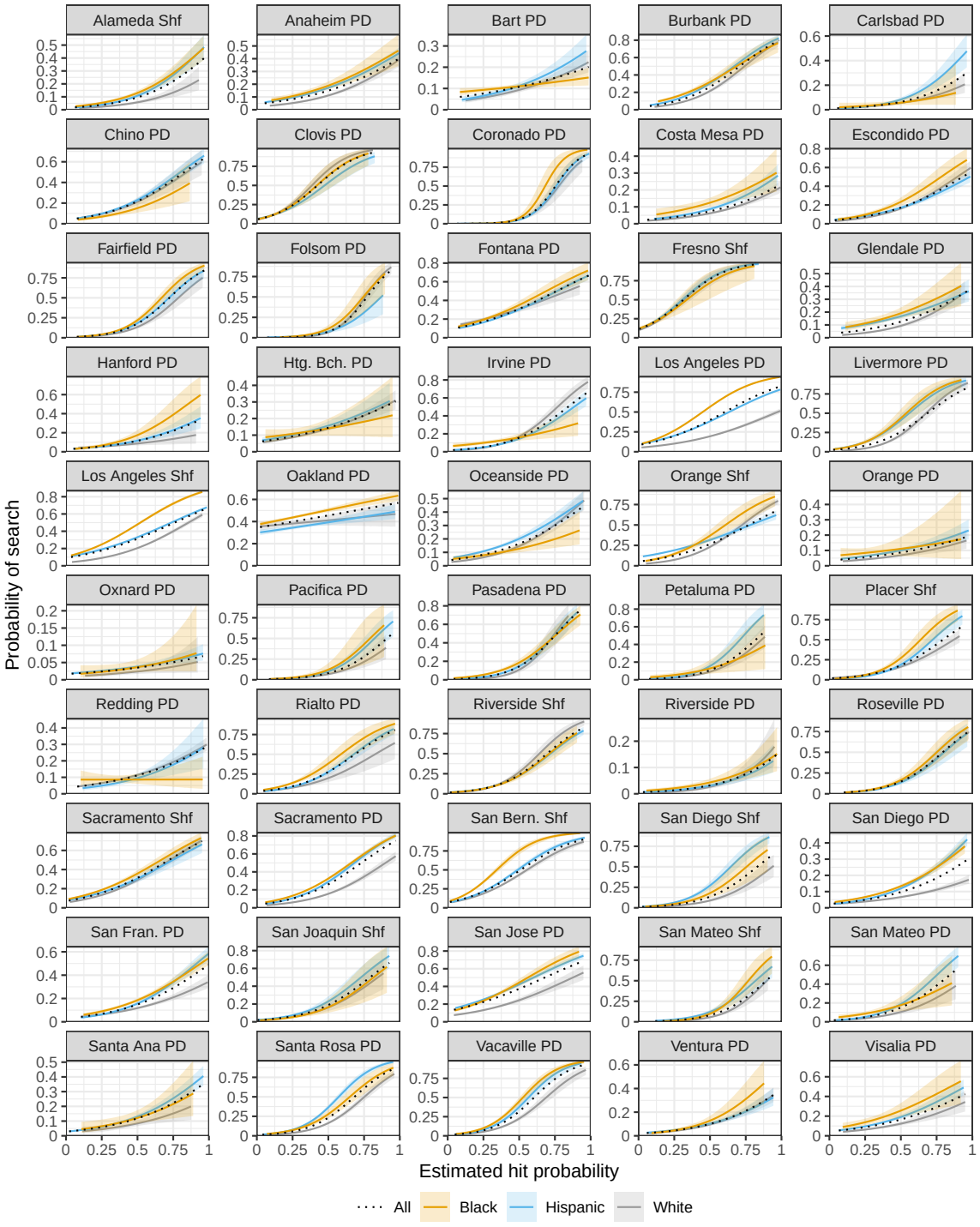

Figure A10: *For each of the 50 largest agencies, observed probability of search as a function of estimated probability of recovering contraband. Lines are fit via logistic regression, with 95% confidence bands. For agencies where the error bands do not overlap, such as the Los Angeles Police Department, the observed difference in discretionary search rates (i.e., adverse impact) cannot be fully explained by the likelihood of recovering contraband from a discretionary search.*

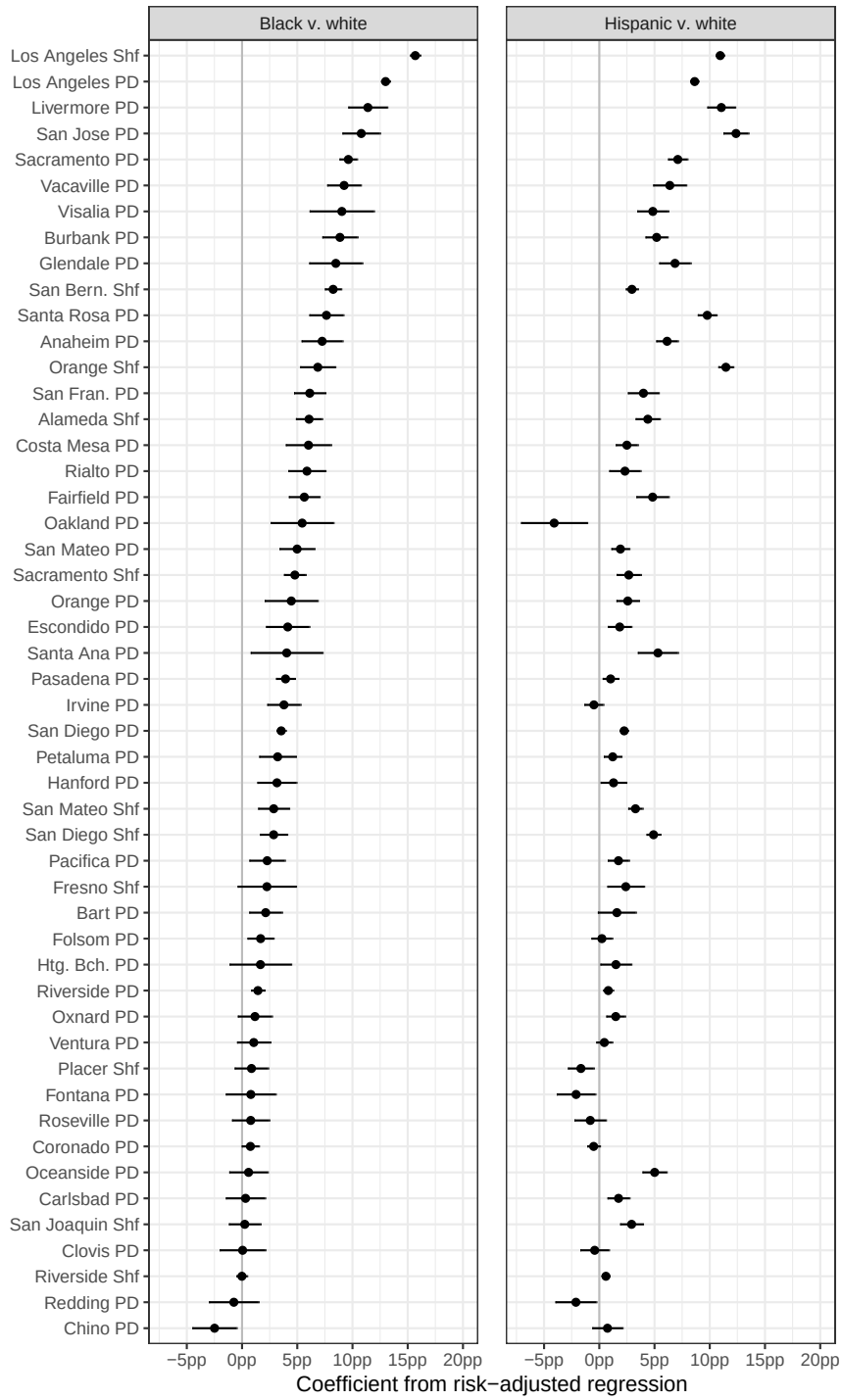

Figure A11: *For each of the 50 largest agencies, race and ethnicity coefficients from a risk-adjusted regression model fit separately to each agency, with 95% confidence intervals. When confidence intervals do not overlap with the 0pp line, the observed differences in search rates across race/ethnicity cannot be fully explained by the estimated risk of recovering contraband.*

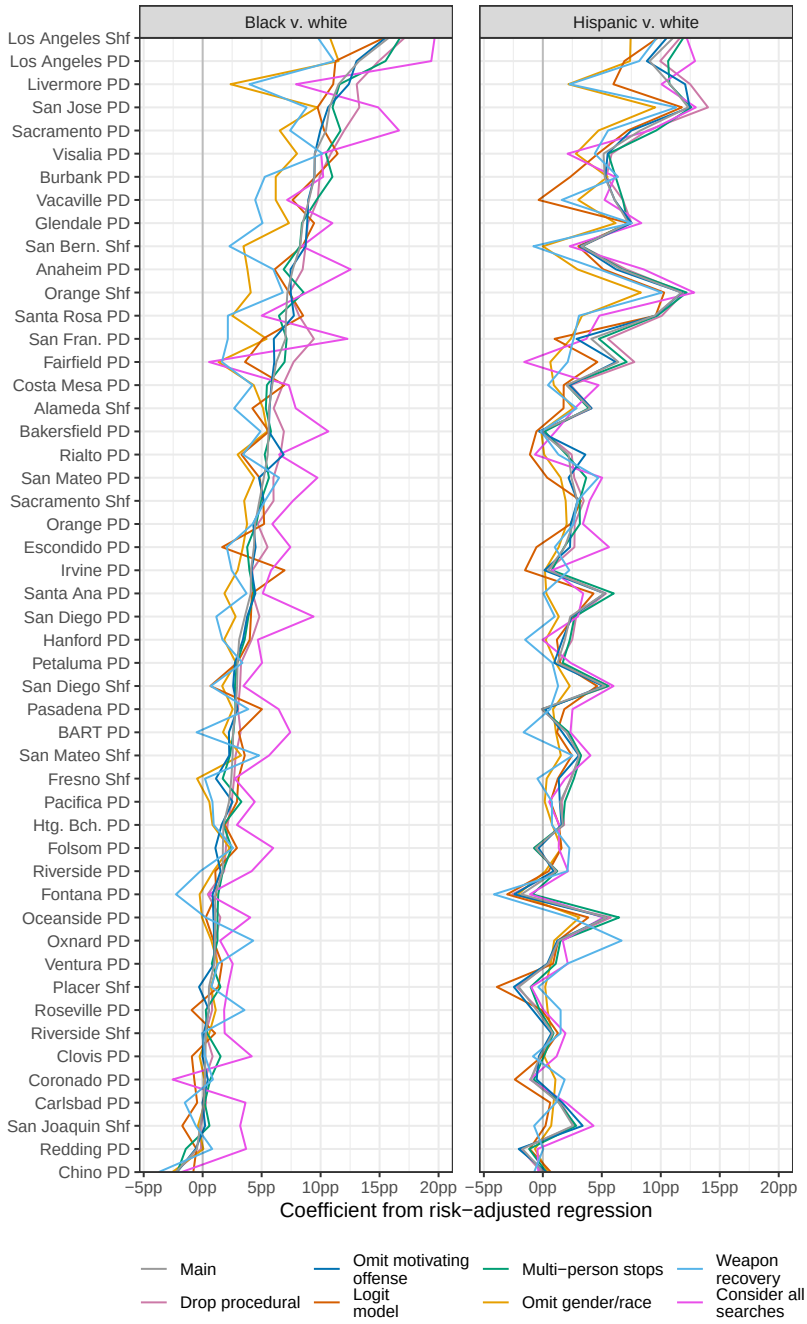

Figure A12: *Robustness checks of the risk-adjusted regression models. "Main" indicates the race and ethnicity coefficients from the unaltered models in Figure A11. "Drop procedural" removes stops resulting in a non-discretionary search. "Omit motivating offense" uses risk models that do not account for the motivating traffic violation or suspected offense that prompted each stop. This model simulates what might happen under moderate omitted variable bias, as this covariate is the third most predictive feature in the random forest models. "Logit model" uses a logistic regression with no interaction terms to fit the risk models, instead of random forests. "Multi-person stops" includes all individuals stopped in multi-person encounters, instead of just the individual recorded first. "Omit gender/race" excludes gender and race from each risk model. "Weapon recovery" uses weapon recovery as the outcome of the risk models, as opposed to using any contraband recovery. "Consider all searches" uses the unaltered search label for non-discretionary searches. Results are qualitatively similar across specifications, though coefficients are somewhat attenuated under the "Omit gender/race" and "Weapon recovery" specifications, and coefficients for Black individuals tend to be larger when non-discretionary search labels are unaltered.*

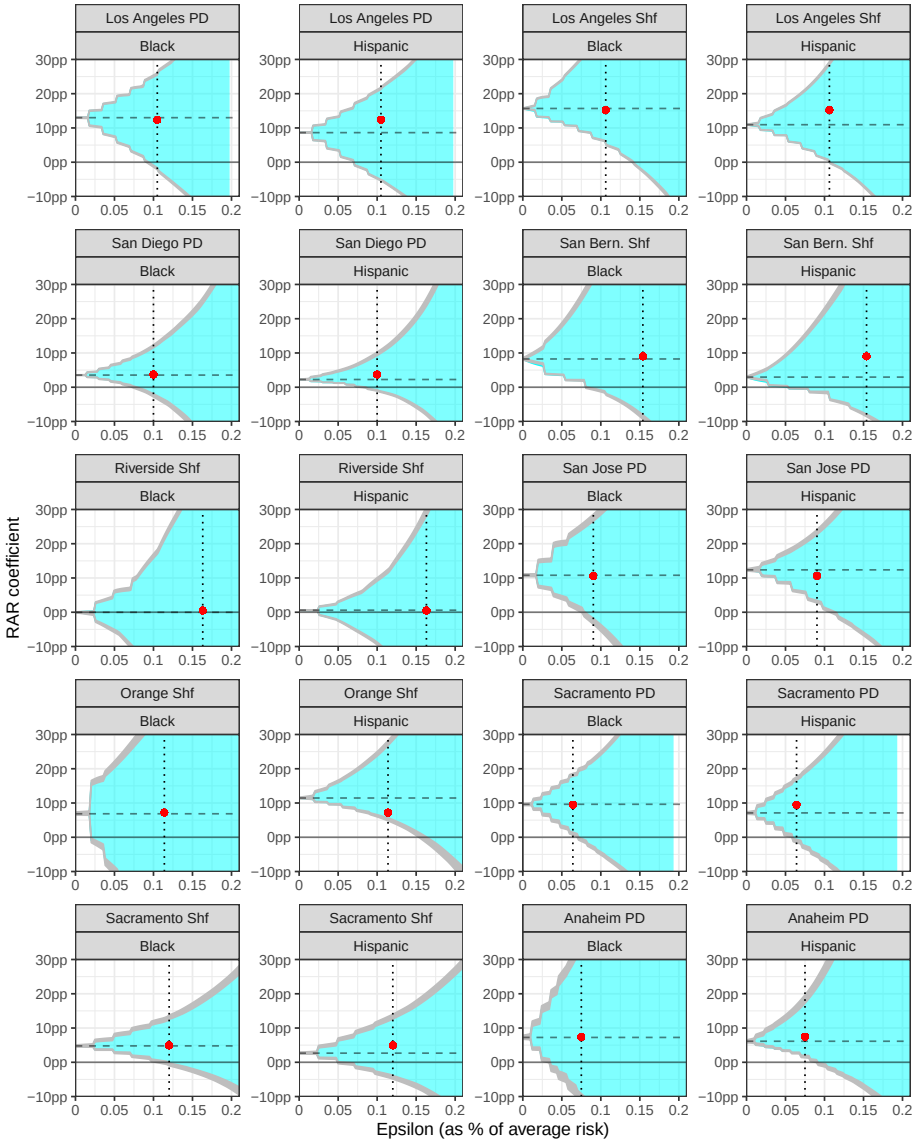

Figure A13: *Sensitivity analysis for the risk-adjusted regression results in Figure 4. The horizontal dashed line shows the risk-adjusted regression coefficient for each race and ethnicity. The x-axis indicates the mean absolute deviation (MAD) between true risk and estimate risk, divided by the overall contraband carrying for the given agency. Jung et al. [2023] refer to this value as ϵ (Epsilon). The blue shaded area indicates the most conservative bounds in the risk-adjusted regression coefficient, conditional on the value of the MAD between true and estimate risk indicated on the x-axis. The gray shaded area shows 95% confidence intervals from a bootstrapping procedure for generating the bounds. The red dot indicates the risk-adjusted regression coefficient from the "Omit motivating offense" specification in Figure A12, which is intended to simulate a degree of moderate omitted variable bias by blinding the model to the traffic violation or suspected offense that prompted the stop. For all agencies, the actual and blinded estimates are quite similar. As a benchmark, the vertical dotted line that intersects the red dot is the observed MAD between risk estimated under the actual model and risk estimated under the blinded model. If the bounds at this benchmark MAD exceed 0pp, then the results are robust to any type of confounding that results in the same MAD. The results for stopped Black individuals are approximately robust to moderate confounding for the Los Angeles PD, the Los Angeles County Sheriff, the Sacramento PD, and the Sacramento County Sheriff. For Hispanic individuals, the results are approximately robust for the Los Angeles County Sheriff, the San Diego PD, the San Jose PD, the Orange County Sheriff, the Sacramento PD, and the Anaheim PD.*

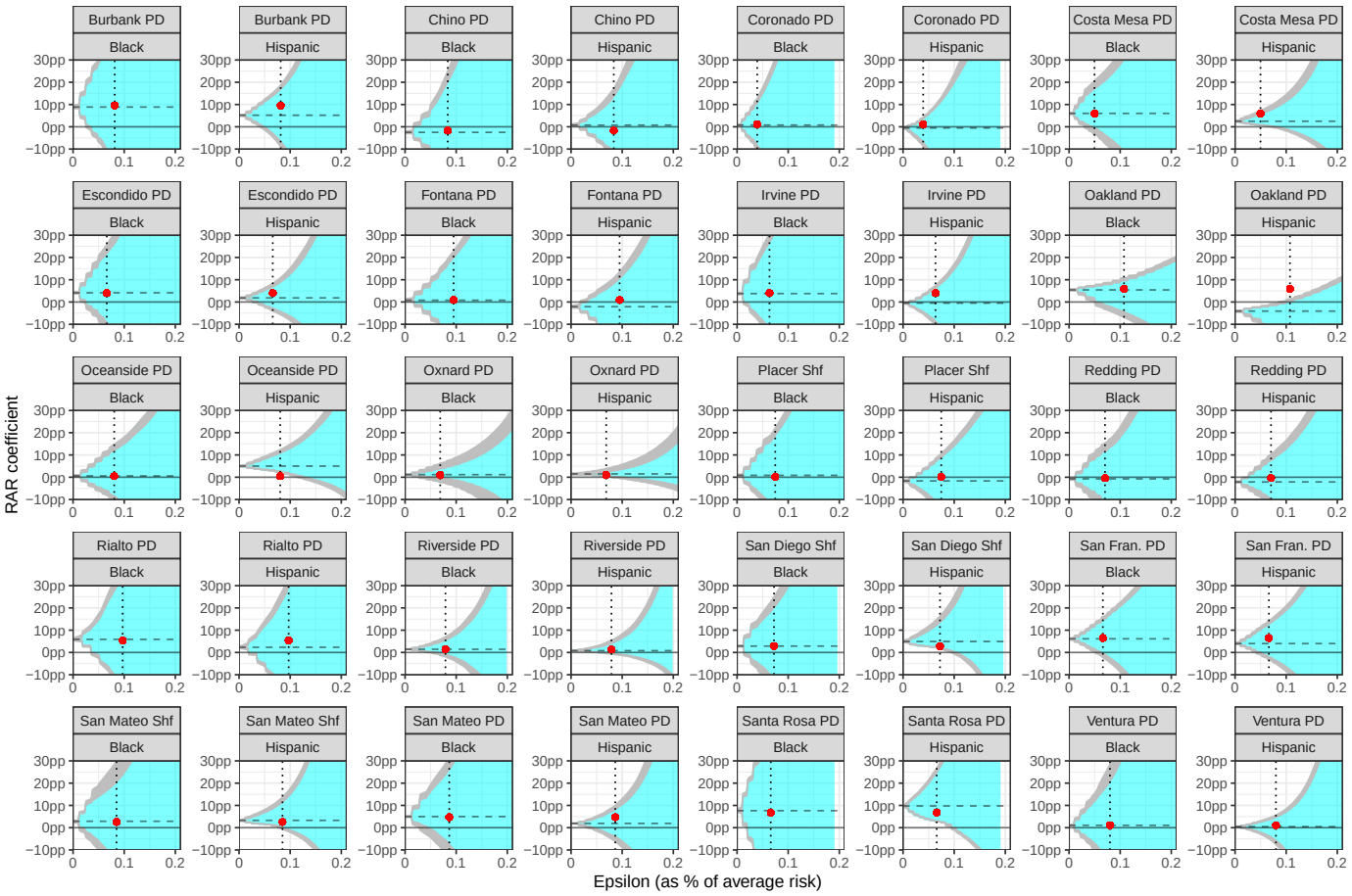

Figure A14: *Sensitivity analysis for the 11th through 30th largest agencies, using the same methods as Figure A13.*

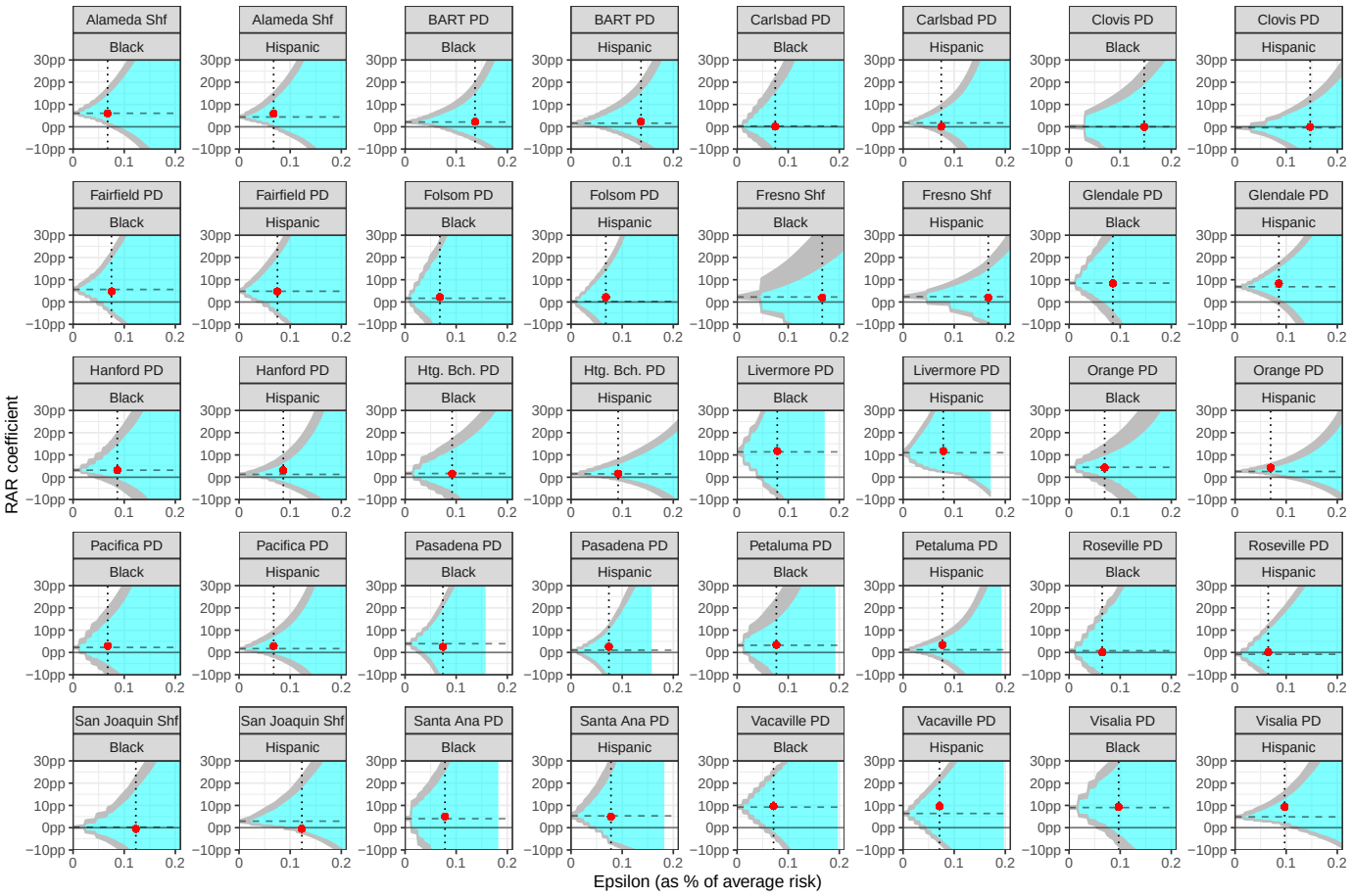

Figure A15: *Sensitivity analysis for the 31st through 50th largest agencies, using the same methods as Figure A13.*

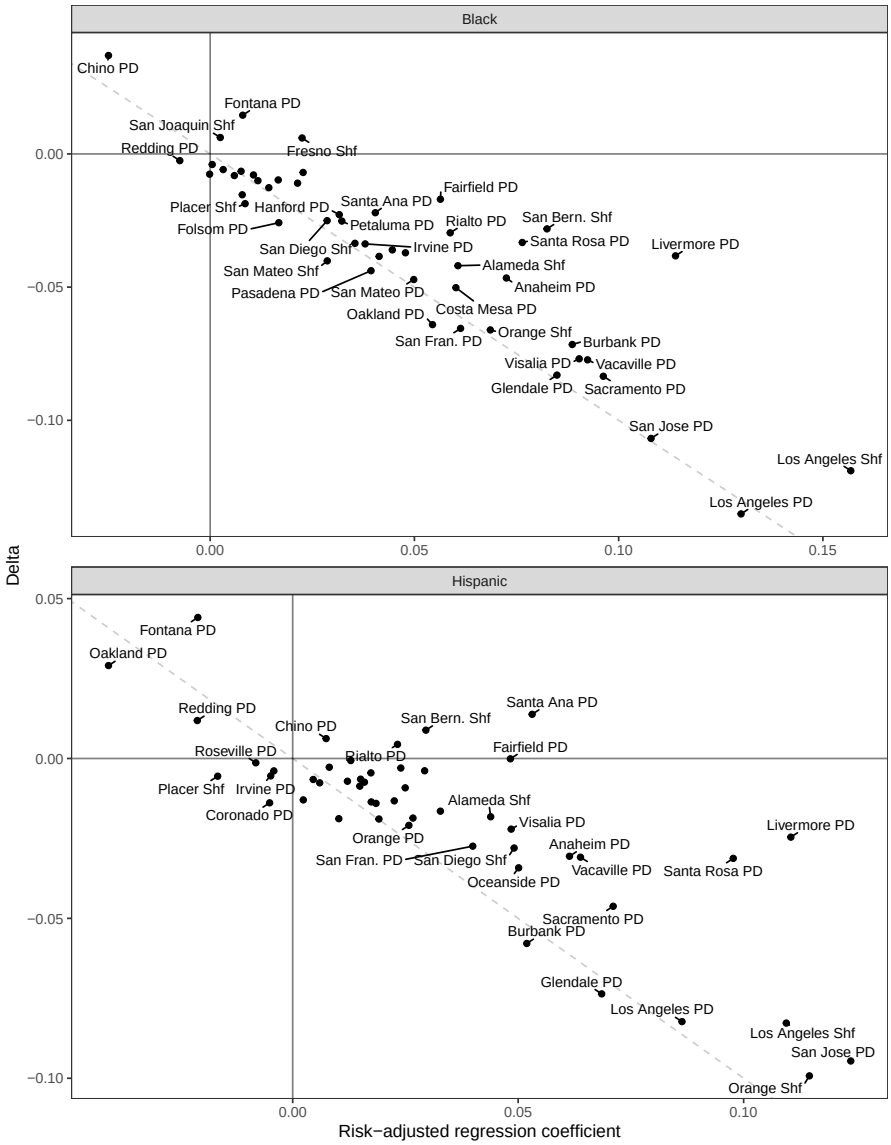

Figure A16: *For each of the 50 largest agencies, comparison of the Δ (Delta) measure of discrimination proposed by Arnold et al. [2021] to the corresponding estimate from risk-adjusted regression. Across jurisdictions, the estimates tend to be similar in magnitude, though there are notable deviations. For example, the Δ measure is close to zero for Hispanic drivers stopped by the Fairfield Police Department, yet risk-adjusted regression suggests a significant difference in risk-adjusted search rates. This discrepancy is possibly a result of the inframarginality concerns raised in the main text. Indeed, the distributions of estimated risk among Hispanic and white drivers stopped by the Fairfield Police Department are quite different (See Figure A17). Figure A18 is a zoomed-in version of this plot with labels for points closer to the origin.*

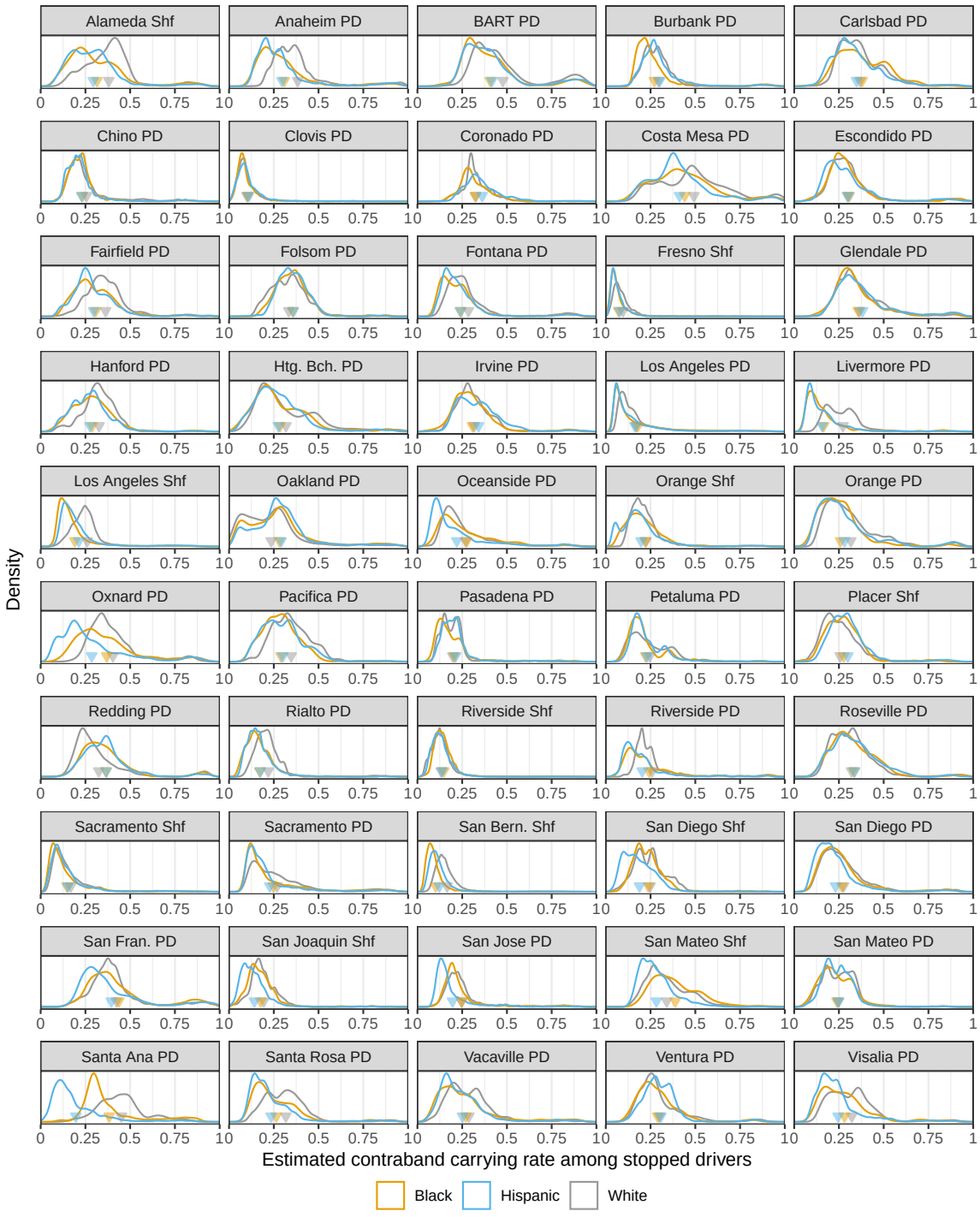

Figure A17: *Distribution of estimated risk by race and ethnicity across the 50 largest agencies. While the distributions are quite similar for certain agencies (e.g., the Sacramento Police Department and County Sheriff), they are markedly different for other agencies (e.g., the Santa Ana Police Department). When underlying distributions of risk differ across groups, inframarginal statistics, such as differences in error rates, may falsely suggest or refute discrimination.*

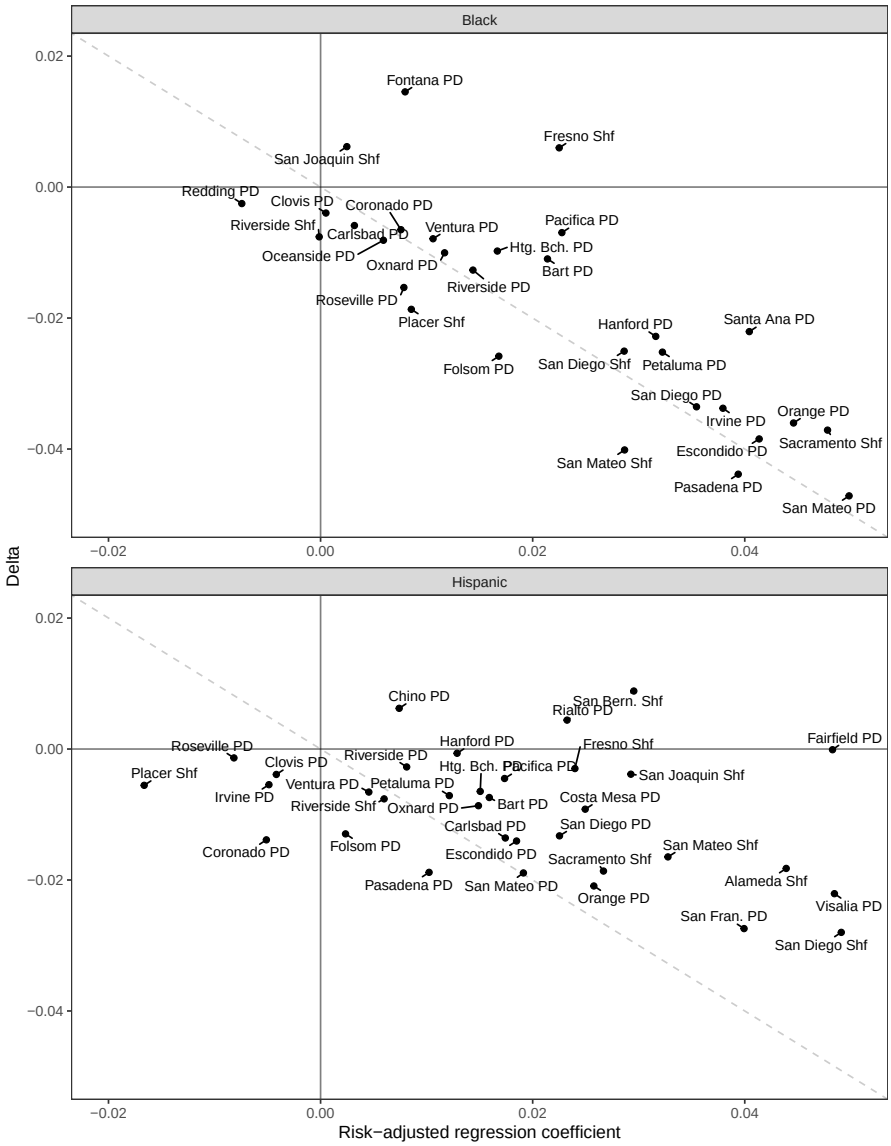

Figure A18: *Zoomed in version of Figure A16 for identifying agencies close to the origin.*

Constructing the simple rule

To construct the simple rule risk model, we begin with the same set of covariates as those listed in Table A2. Figure A7 shows that, across jurisdictions, plain view contraband and evidence of a crime are the most predictive factors of contraband recovery from a search, with predictive feature importance dropping off quickly. Risk varies substantially with agency. As such, we fit agency-specific logistic regression models to estimate the likelihood of recovering contraband using the city, the traffic violation or suspected offense than prompted the stop, and the two most predictive factors: whether the search was prompted by contraband in plain view, and whether the search was prompted by evidence of a crime. We fit this model only on individuals who were searched at the discretion of the officer, as the contraband recovery outcome is unknown for individuals who are not searched.

With this fitted model in hand, we multiply the fitted coefficients of the two key factors by 10 to put the coefficients on an approximate integer scale, and then round the two coefficients to the nearest integer. Using just these two rounded coefficients, we calculate a risk score for each searched individual. For example, if the rounded coefficients are 10 for plain view contraband and 15 for evidence of a crime, a stopped individual with plain view contraband but no evidence of a crime would receive a score of 10. Finally, we fit an additional logistic regression model that predicts contraband recovery using this score, the city, and the motivating offense that motivated the stop. This model is again fit just to searched individuals. This final fit provides the optimal coefficients for each city and motivating offense.

To operationalize the simple rule, each agency could select a risk threshold above which officers are obligated to conduct a discretionary search (e.g., search if there is at least a 20% chance of recovering contraband). Using the final fitted model, each agency would determine, for each combination of city and motivating offense, the minimum score needed for the estimated risk of contraband recovery to exceed the desired threshold (e.g., 20% risk might correspond to a score of 12 if the stop occurs in Santa Clara and the stop was prompted by a speeding driver). Before initiating a stop, officers could report the city and motivating offense to obtain a risk threshold. Once the stop begins, officers could quickly calculate the individual's risk score using the two factors, and then conduct a search if the score exceeds the known threshold.

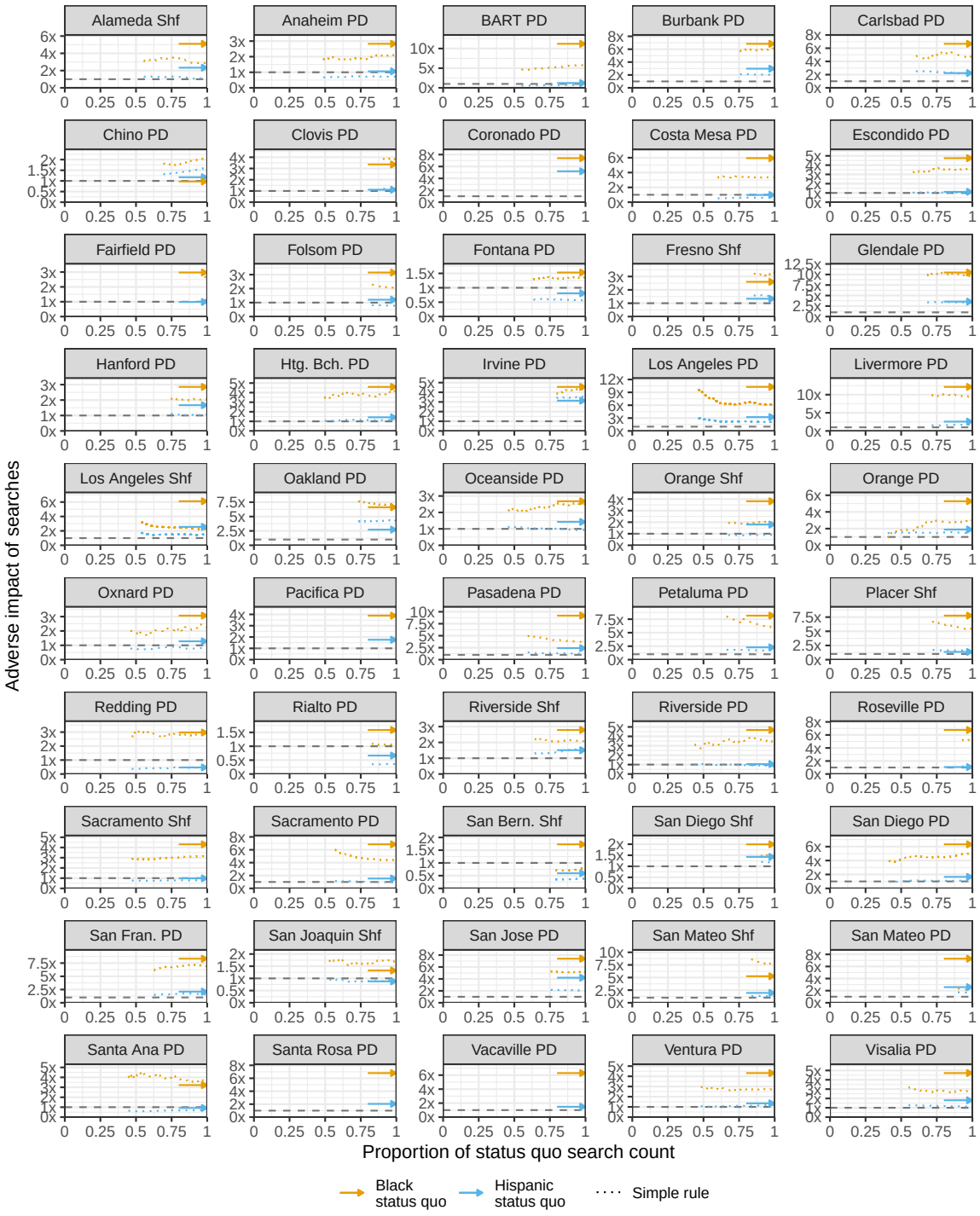

Figure A19: *Expanded version of Figure 5 for all 50 agencies.*

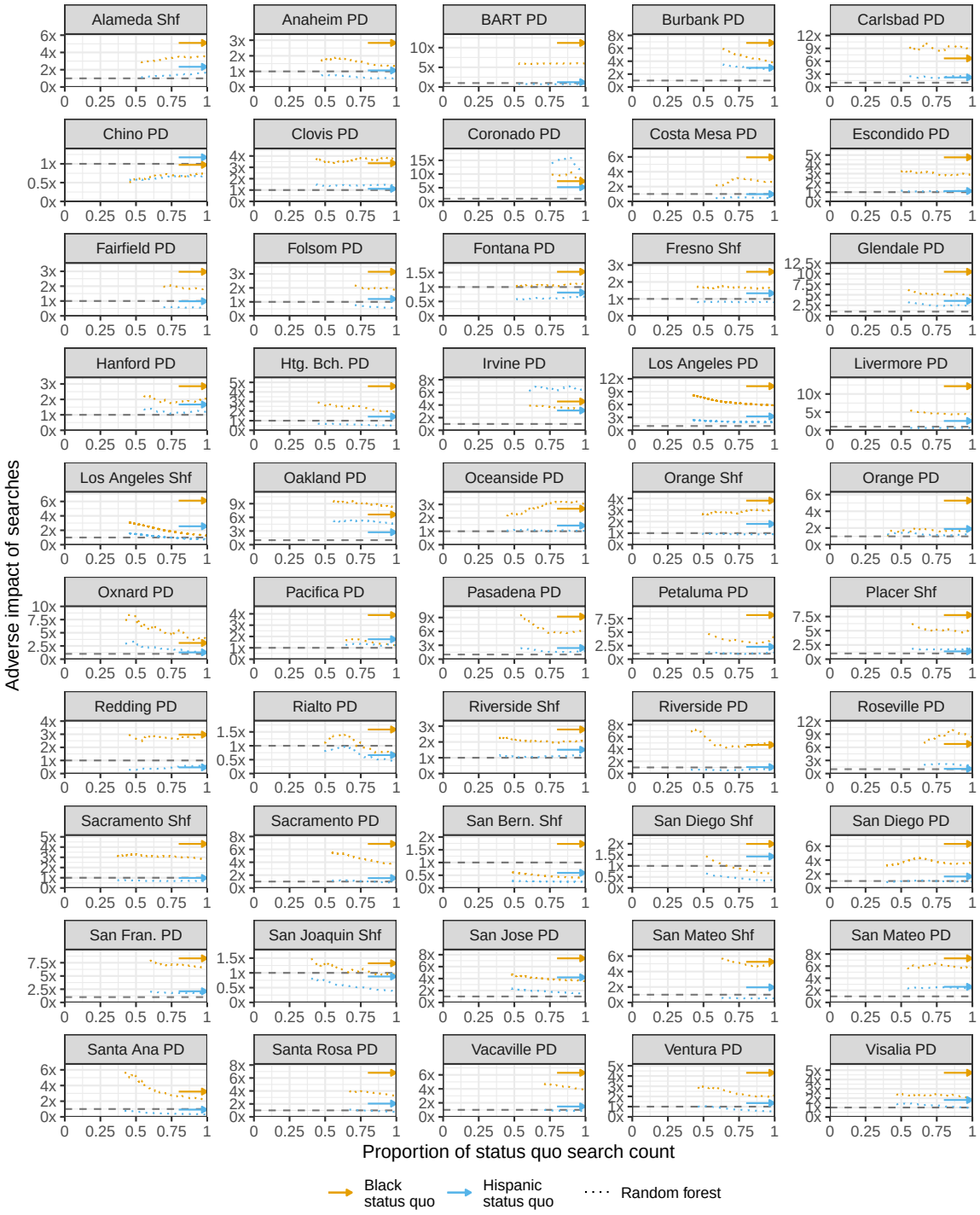

Figure A20: *Expanded version of Figure 5 for all 50 agencies. Instead of using a simple rule risk model, these plots use a random forest risk model to generate risk estimates.*

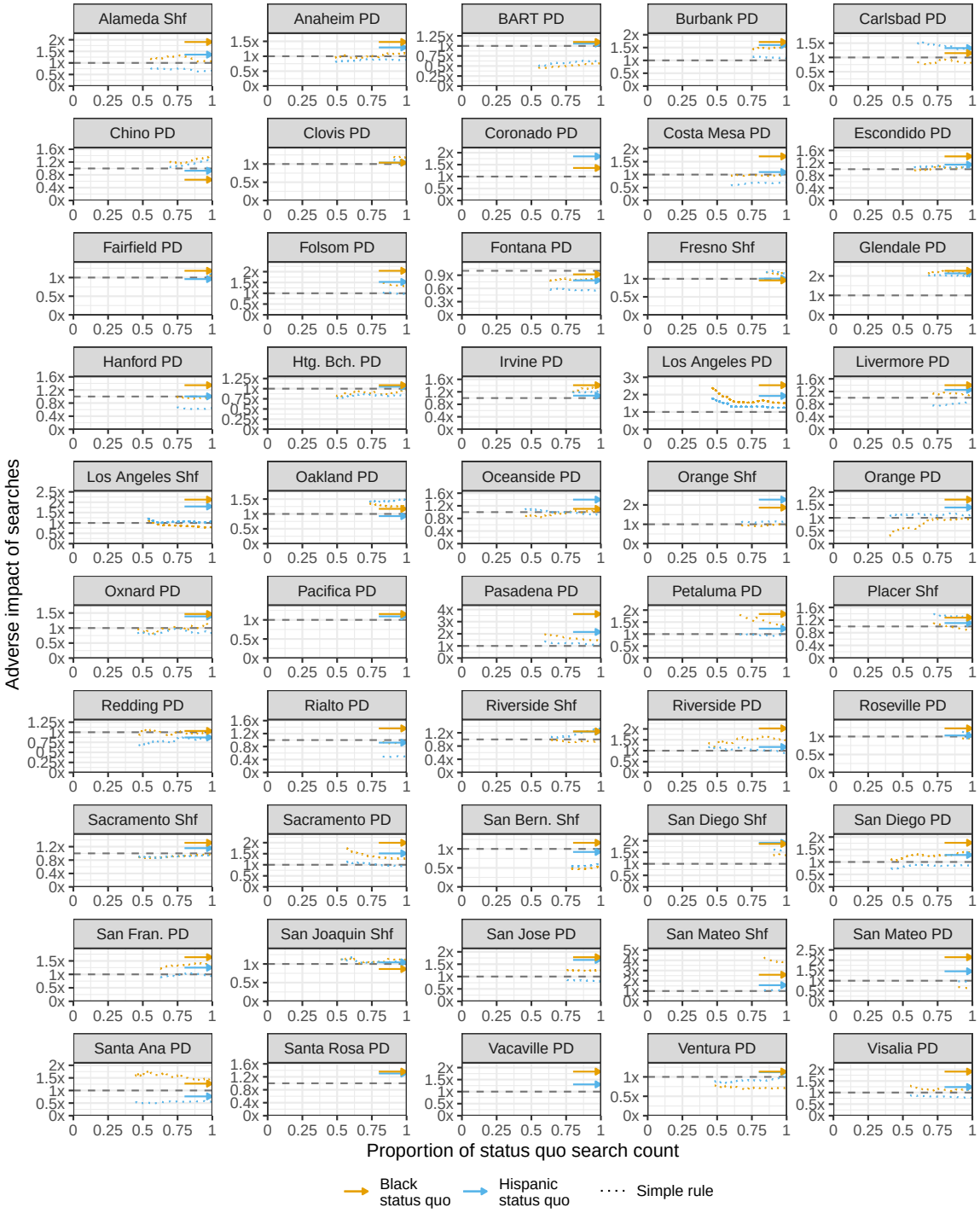

Figure A21: *Expanded version of Figure 5 for all 50 agencies. Instead of measuring adverse impact with a population-level benchmark, these plots use a stop-level benchmark (see Figure A1).*