Measuring discrimination using natural experiments *

Nikhil Rao† James Reeves‡

November 14, 2025

Abstract

Disparities between social groups (e.g., gender, race) exist at crucial decision points in many contexts, from hiring to policing. These disparities are difficult to interpret when unobservable factors vary by group. To overcome this challenge, we use a binary instrumental variable (IV) strategy to measure discrimination after adjusting for group differences in unobserved potential outcomes, with extensions to adjust for treatment effects or multiple potential outcomes instead. Assumptions on selection behavior recover the distribution of potential outcomes for each group, which we directly condition on to bound or point identify discrimination. This approach expands where researchers can measure such discrimination since existing tools need randomly assigned decision-makers. We use our method to study racial discrimination in misdemeanor prosecution, the most common form of contact with the United States criminal court system. We use a difference-in-difference IV strategy generated by a budget cut that reduced prosecution rates in King County, Washington, but did not affect adjacent counties. Before the cut, we find no evidence of discrimination in prosecution conditional on unobserved potential recidivism. After the cut, minority defendants became less likely to be prosecuted than white defendants with similar potential recidivism. We find suggestive evidence that this pattern is driven by prosecutors dismissing lower quality, resource-intensive cases. Such cases were more prevalent among minority defendants, perhaps due to disparities created in criminal legal decisions before prosecution.

*We thank Ashley Craig, Sara Heller, Sarah Miller, Michael Mueller-Smith, Benjamin Scuderi, Kevin Stange, Mel Stephens, and Basit Zafar for their comments and guidance. We also thank Lea Bart, Micah Y. Baum, Jordy Berne, Charlie Brown, Elisa Facchetti, Jamie Fogel, Benjamin Hansen, Christopher Hollrah, Emily Horton, Peter Hull, Brian Jacob, Amanda Kowalski, Steven Mello, Emir Murathanoğlu, Tyler Radler, Katherine Richard, Roman Rivera, Brock Rowberry, Damián Vergara, Iris Vrioni, and numerous seminar participants for helpful suggestions. We are grateful to Kevin Cottingham at the Washington Administrative Office of the Courts for assistance with the court data and to Jordy Berne, Brian Jacob, and Christina Weiland for their help in working with the 'Read by Grade 3' project data. Rao gratefully acknowledges financial support from the Poverty Solutions Doctoral GSRA award, the Rackham One Term Dissertation Fellowship, and the Arnold Ventures Post-doctoral Fellowship on the Economics of Crime, awarded through the National Bureau of Economic Research. Reeves thanks Arnold Ventures for financial support.

†National Bureau of Economic Research: nikhrao@umich.edu

‡University of Colorado Denver: james.reeves@ucdenver.edu.

1 Introduction

Disparities between social groups (e.g., race, gender, socio-economic status) are common in many contexts, e.g., employment, housing, and the criminal legal system. To understand whether such disparities are the result of discrimination, researchers often seek to compare how similar individuals from two groups are being treated.[1] This comparison is difficult to make in practice if there are important but unobserved group differences (e.g., Aigner and Cain, 1977; Charles and Guryan, 2011). For example, in the canonical context of hiring discrimination, it is challenging to measure the hiring gap between equally productive people from two groups, because productivity is only observed for hired workers.

An extensive literature has focused on overcoming similar challenges by operationalizing the Becker (1957) 'outcome' test, using theoretical restrictions and quasi-experimental tools (Knowles, Persico, and Todd, 2001; Anwar and Fang, 2006; Arnold, Dobbie, and Yang, 2018; Marx, 2022; Canay, Mogstad, and Mountjoy, 2024; Bharadwaj, Deb, and Renou, 2024). These tests detect taste-based discrimination (Arrow, 1973) at the margin of a decision, but do not quantify accurate statistical discrimination (Phelps, 1972) or discrimination away from the margin, regardless of its source (Hull, 2021). Recent work addresses these concerns by using random assignment to decision-makers to extrapolate unobserved outcomes (e.g., productivity of workers not hired) and directly adjust for these unobservables when measuring discrimination (Arnold, Dobbie, and Hull, 2022). While this approach captures discrimination whether it is taste-based, statistical, or due to incorrect beliefs, it cannot be applied when individuals are not randomly assigned to decision-makers.

In this paper, we expand where researchers can credibly measure discrimination. We use natural experiments that yield a binary instrumental variable (IV) framework (Angrist, Imbens, and Rubin, 1996) to measure average discrimination in a treatment decision among individuals with similar potential outcomes or treatment effects. We map commonly used quasi-experimental techniques, e.g., regression discontinuity (RD) or difference-in-difference (DiD), to the binary IV framework and use assumptions on selection behavior to recover the distribution of unobserved outcomes for each group. We directly condition on these unobservables to either bound or point identify discrimination that arises from multiple sources.

We implement our approach to document novel evidence on racial discrimination in the context of misdemeanor adjudication, the most common form of contact with the criminal court system. Roughly 4,000 misdemeanor cases are filed annually per 100,000 people in the United States, a rate that is triple the felony filing rate (Stevenson and Mayson, 2018). These large caseload sizes can create incentives to process cases quickly with limited oversight, potentially exacerbating disparities and discrimination (Gershowitz and Killinger, 2011). We study racial discrimination in misdemeanor prosecution in King County, Washington (Seattle metropolitan and suburban areas). Cases are not randomly assigned to prosecutors in this setting, rendering existing discrimination

[1] Prior work documents disparities and discrimination in hiring (e.g., Hellerstein, Neumark, and Troske, 1999, 2002; Goldin and Rouse, 2000; Bertrand and Mullainathan, 2004), housing (e.g., Ross, 2005; Ross et al., 2008), and the criminal legal system (e.g., Abrams, Bertrand, and Mullainathan, 2012; Anwar, Bayer, and Hjalmarsson, 2012).

measurement approaches infeasible. Instead, we construct a binary IV using a DiD strategy generated by an unexpected budget cut that affected prosecutors in King County but not in adjacent counties. Our data consist of all cases that prosecutors accept from law enforcement and we examine the decision to continue to pursue an accepted case. Throughout the paper, we use the term 'prosecuted' if a case was pursued and all charges were not dismissed.

To outline the details of our approach, consider a potential outcomes (Rubin, 1974) framing, where $D_i = 1$ if an individual is prosecuted (i.e., treated) and $D_i = 0$ if dismissed. We define racial discrimination as the racial gap in prosecution rates between individuals who would have the same potential re-offence outcome, motivated by prior work suggesting that potential re-offending is an important consideration in misdemeanor prosecution (Agan, Doleac, and Harvey, 2023). We focus our exposition on the racial gap in prosecution conditional on the re-offence outcome if prosecuted, $Y_i(1)$, but later discuss extensions to condition on i) the re-offence outcome if dismissed, ii) the treatment effect of prosecution, and iii) two potential outcomes simultaneously (e.g., re-offence risk and case quality).[2] In each of these cases, if the distributions of the unobservable factors vary by race, then the raw prosecution disparity could be driven by the unobservable differences and would not align with our definition of discrimination. For simplicity, let the treated potential outcome be an indicator for re-offending if prosecuted. With binary $Y_i(1)$, the distribution of potential recidivism if prosecuted is the share of people who would re-offend if everyone were prosecuted—the "average prosecuted outcome".

If we could observe the distribution of prosecuted outcomes for each racial group, we would quantify discrimination conditional on the prosecuted outcome in three steps. First, for each racial group and value of the prosecuted outcome, we rescale the observed race-specific prosecution rates by the ratio of the share of prosecuted defendants to the share of the population with that outcome value (Arnold, Dobbie, and Hull, 2022). Intuitively, this ratio reflects how selected the sample of prosecuted defendants is in terms of the prosecuted outcome. Rescaling by this ratio yields race-specific prosecution rates that are conditional on the prosecuted outcome. Second, we use these prosecution rates to construct racial gaps in prosecution for each value of the prosecuted outcome. That is, we construct racial gaps for people who: a) would re-offend if prosecuted ($Y_i(1) = 1$) and b) would not re-offend if prosecuted ($Y_i(1) = 0$). Third, the weighted average of these two racial gaps yields average discrimination in prosecution conditional on the prosecuted outcome.

To overcome the challenge of not observing the average prosecuted outcome for each racial group, we estimate them using a binary IV that shifts treatment rates—prosecution rates in this example. First, we use the IV to estimate how potential outcomes (e.g., recidivism if prosecuted) vary across defendants who would always be prosecuted ('always takers') and marginal defendants prosecuted due to the IV ('compliers') within each racial group (Imbens and Rubin, 1997). However, we do not observe the outcomes that dismissed defendants ('never takers') would have experienced

[2]Recent work critiques discrimination definitions that condition on potential outcomes and suggests alternative definitions that may be more or less appropriate depending on the context (Grossman, Nyarko, and Goel, 2024; Bushway et al., 2025). While the focus of this paper is expanding where researchers can measure discrimination conditional on potential outcomes, we discuss these critiques in Section 2.1.

if they were prosecuted instead. We extrapolate their prosecuted outcomes by using behavioral assumptions from the marginal treatment effects literature that restrict the relationship between average potential outcomes across 'always takers', 'compliers', and 'never takers' (Brinch, Mogstad, and Wiswall, 2017; Mogstad, Santos, and Torgovitsky, 2018; Kowalski, 2023a,b). Depending on the strength of the assumptions, such extrapolations recover either bounds or point estimates of the unobserved average prosecuted outcomes for each racial group.[3,4,5]

Our approach also applies when a natural experiment generates a conditionally-random binary IV. In particular, we discuss implementation with a DiD, a setup where time might independently influence treatment status and potential outcomes, making it difficult to identify the proportions and outcomes of always takers, compliers, and never takers. We show how restrictions on the effect of time and an assumption of parallel trends in potential outcomes permit identification of these key quantities that are crucial for our approach.

Taking our approach to the data, we use administrative court records from Washington State to bound racial gaps in misdemeanor prosecution among individuals who are equally likely to commit a new offence if prosecuted.[6] Using the shifts in prosecution rates and re-offence outcomes generated by the budget cut, we find racial differences in the unobserved prosecuted outcome. Before the budget cut, 24–29% of white defendants would commit a new offence if prosecuted, while the same is true for 32–37% of minority defendants. These stark differences imply that discrimination estimates that do not account for differences in unobserved prosecuted outcomes, e.g., covariate-adjusted prosecution gaps, will be biased.

We use the estimates of the average outcomes if everyone were prosecuted to bound the racial gap in prosecution before and after the budget cut, conditional on the prosecuted outcome. Before the budget cut, we cannot reject the null of no racial gap in prosecution rates. This pattern changes after the cut, which sharply reduced prosecution rates—minority defendants became 1.3–4 p.p. (1.7–5.3% of post-period white prosecution rates) less likely than white defendants to be prosecuted. The gap implied by our bounds is lower than and excludes the covariate-adjusted disparity (4.5 p.p.). Thus, we find evidence of unwarranted disparities in prosecution after the budget cut that favored minority defendants, but less than what naïve estimates would have concluded.

Next, we investigate potential reasons for the relatively lower minority prosecution rate after the budget cut. One explanation might be that: i) resource constraints after the budget cut made prosecutors less likely to pursue low quality cases, and ii) minority cases were supported by

[3]Compared to other approaches to bound unobserved outcomes (e.g., Manski, 1989; Jordan, 2024), our approach will provide tighter bounds in practice when average outcomes for compliers and always takers (in the case of treated potential outcomes) are quite different, especially when a large share of the population has unobserved outcomes.

[4]To simultaneously condition on treated and untreated outcomes or two distinct potential outcomes, we first use this approach to extrapolate population averages for each potential outcome of interest. Then, we bound the covariance between the potential outcomes using information from the extrapolations and observed means, following Hössjer and Sjölander (2022). See Appendix B.1 for details and the mapping to conditioning on treatment effects.

[5]While we estimate bounds in our empirical application, Appendix B.2 illustrates the point-identification approach using a brief empirical example studying racial discrimination in incarceration decisions in Texas.

[6]Since racial differences in sentencing might complicate the interpretation of discrimination conditional on re-offence outcomes if prosecuted, we also present exercises conditioning on the re-offence outcome if dismissed, which compare defendants who are similar in the baseline likelihood of re-offending.

weaker evidence due to discrimination in pre-prosecution decisions, e.g., police arrest or prosecutor charging decisions (Goncalves and Mello, 2021; Owens and Ba, 2021; Agan, 2024). If this is the case, we should expect more pronounced white–minority gaps after the budget cut among low quality cases. We find that the post-cut racial gap is driven by low quality cases, both when using a data-driven procedure to define low quality cases (based on historical sentencing success rates) and when simultaneously conditioning on two potential outcomes: recidivism and case success if prosecuted. We interpret these patterns as suggestive evidence that prosecutors shifted their focus to high quality cases after the budget cut, offsetting disparities potentially created in stages of the criminal legal system before the prosecution decision. These patterns mirror recent work on prosecutorial discretion (Harrington and Shaffer, 2023; Jordan, 2024) and highlight the importance of considering the criminal legal system as a multi-stage institution (Harrington and Shaffer, 2024).

As a secondary application, we also discuss how to combine our approach with the local variation generated by an RD to measure discrimination conditional on potential outcomes at an RD cut-off and away from it. Appendix B.3 empirically illustrates this point, using an RD design to measure socio-economic discrimination in the decision to advance public school students to the next grade.

This paper makes several contributions. First, we expand where researchers can measure discrimination conditional on key unobserved factors. Since our approach only requires a binary IV and assumptions from the marginal treatment effects literature (Brinch, Mogstad, and Wiswall, 2017; Mogstad, Santos, and Torgovitsky, 2018; Kowalski, 2023a,b), it is especially useful when individuals are not randomly assigned to decision-makers (as in our empirical application) or if researchers do not even observe decision-makers in the data, conditions that render existing approaches inapplicable. For example, since students or employees are rarely randomly assigned to teachers or managers, researchers could use our approach to study discrimination in student suspensions or worker promotions. Our method also lets researchers condition on treatment effects or combinations of distinct potential outcomes, which can be useful when it is appropriate to consider more than one unobserved factor when studying unwarranted disparities. Beyond discrimination, our approach broadens where researchers can estimate decision rates conditional on potential outcomes, which are useful for studying decision-maker accuracy (e.g., Angelova, Dobbie, and Yang, 2023; Kleinberg et al., 2018) and the equity and efficiency of policy targeting (e.g., Rose, 2021). As a final methodological point, we add to the literature mapping DiD to IV (De Chaisemartin and D'Haultfœuille, 2018) by showing how to use DiD variation to estimate average potential outcomes.

Second, our empirical analysis presents the first evidence of racial discrimination in misdemeanor prosecution that documents and directly accounts for unobservable racial differences. Although recent work has examined disparities in misdemeanor prosecution (Kutateladze and Andiloro, 2014; Sloan, 2022; Amaral, Ouss, and Ozier, 2025; Jordan, 2025), most of the prior literature on disparities in prosecutorial decisions has focused on felony prosecution (e.g., Spohn, Gruhl, and Welch, 1987; Davis, 2014; Rehavi and Starr, 2014; Yuan and Cooper, 2022; Harrington and Shaffer, 2023; Tuttle, 2023; Agan, 2024; Jordan, 2024). In either case, much of this evidence is either descriptive or from contexts where unobservable differences between the comparison groups are minimal.

2 Estimating discrimination using a binary instrument

In this section we show how to use a binary instrumental variable (IV) to measure discrimination conditional on potential outcomes. First, we formalize the definition of discrimination that we are interested in measuring. Second, we highlight which components of the definition are observed in the data and which are not. Third, we show how to use insights from the IV and marginal treatment effects literatures to estimate the unobserved components, and discuss practical details related to implementing our approach.

2.1 Discrimination estimands of interest

We begin with a general potential outcomes framework in which individuals are chosen for a binary treatment, D_i, as a function of the unobservable potential outcomes, where $Y_i(D_i = 1)$ is the treated outcome and $Y_i(D_i = 0)$ is the untreated outcome (Rubin, 1974; Imbens and Angrist, 1994). Each potential outcome is only observed if the associated treatment state is realized. These potential outcomes may be continuous, discrete, or binary. Individuals belong to one of two groups, denoted by $R_i \in \{r_1, r_2\}$ and the distribution of potential outcomes may differ across groups.

To simplify exposition, we focus our discussion on measuring differential treatment between individuals conditional on the treated potential outcome, although we discuss extensions to i) condition on the untreated potential outcome, ii) condition on the treatment effect, and iii) simultaneously condition on two treated/untreated potential outcomes in Appendix B.1.[7] While group differences in the distribution of potential outcomes might arise due to discrimination in stages upstream or downstream of the treatment decision at hand, our definition of discrimination that conditions on potential outcomes (or treatment effects) is agnostic about the source of the underlying differences. Rather, we seek to understand how much of any group differences in treatment might be captured by the underlying unobservables. As noted in prior work, such a definition of discrimination encompasses multiple sources of discriminatory behavior (statistical discrimination, animus, and biased beliefs) and is consistent with the legal interpretation of 'disparate impact' (Becker, 1957; Phelps, 1972; Arrow, 1973; Bordalo et al., 2016; Bohren et al., 2019; Arnold, Dobbie, and Hull, 2022).[8,9]

In general, the most appropriate unobservable factor to condition on should be such that any group differences in treatment that remain after accounting for differences in the unobservable can be interpreted as an unwarranted disparity. For example, in the canonical context of hiring

[7] Our framework lets us consider quantifying discrimination conditional on any potential outcome as well as treatment effects since it is a general version of the framework described in Arnold, Dobbie, and Hull (2022). In their setting of bail release decisions, individuals vary in a latent unobservable Y^*, which is only observed among treated individuals, making the treated potential outcome synonymous with the treatment effect.

[8] This definition does not require individuals to be identical in terms of all non-race characteristics, as in Canay, Mogstad, and Mountjoy (2024). We think of differences in observed covariates, even conditional on potential outcomes, as possible drivers of discrimination, and investigate them in our empirical application.

[9] Using the pretrial detention context as a motivating example, Grossman, Nyarko, and Goel (2024) and Bushway et al. (2025) argue that such definitions are unsuitable for measuring discrimination, preferring definitions that compare individuals with similar predicted misconduct. While such definitions may be attractive from a legal standpoint in certain contexts, they may not always compare unobservably similar individuals, especially if prediction quality varies by group (e.g., race).

discrimination, any group differences in hiring rates that remain after accounting for differences in worker productivity upon being hired (treated) could be interpreted as unwarranted. In other contexts it may instead be more appropriate to condition on the untreated outcome or the treatment effect. Consider the decision to nominate students for advanced educational programs. Researchers studying discrimination in such a setting may want to quantify group differences in nominations between students who would: i) do equally well without the program ($Y_i(0)$) or ii) have equal gains from the program ($Y_i(1) - Y_i(0)$). The exact choice of potential outcome will typically depend on a combination of the empirical, theoretical, and normative features of the particular context.[10]

Definition 1. Differential treatment conditional on treated potential outcome $Y_i(1)$

$$\triangle_y \equiv E[D_i|R_i = r_1, Y_i(1) = y] - E[D_i|R_i = r_2, Y_i(1) = y]$$

Focusing on discrimination conditional on the treated potential outcome, the estimand described in Definition 1 is generally difficult to estimate empirically. Since treated outcomes are only observed among treated individuals, it is not feasible to directly condition on $Y_i(1)$. The simple approach of computing the raw gap in treatment rates by group will differ from Definition 1 if treatment decisions are a function of $Y_i(1)$ and if the distribution of $Y_i(1)$ varies by group. An alternative approach of computing the treatment gap conditional on a set of observed covariates can also introduce bias if the covariates themselves are generated due to some discriminatory behavior, while also altering the interpretation of the test. Instead of measuring how differently people with the same potential outcome are treated, conditioning on covariates yields a narrower test measuring how much two individuals with identical **observables** are treated differently, generating 'included variables bias' (Ayres, 2010). Even if such bias is small, controlling for covariates will typically not recover the quantity in Definition 1 unless the covariate used is perfectly correlated with $Y_i(1)$.

In practice, groups often differ in terms of unobserved potential outcomes in many settings where measuring discrimination is of interest. For example, discrimination by police could generate cross-race differences in the potential recidivism outcomes among arrested individuals, in turn making it challenging to interpret disparities in prosecution as due to discrimination. As another example, differential access to educational inputs by socio-economic status might generate group differences in student skills, making it difficult to conclude the extent to which any disparities in educational decisions, e.g., student promotion, are due to discrimination.

We next demonstrate how to use a natural experiment to identify the estimand in Definition 1. Consider an intervention that generates a binary instrument, Z, where $Z \in \{0, 1\}$ denotes periods before and after an intervention. Assume Z satisfies the usual instrumental variables (IV) assumptions of relevance, independence, exclusion, and monotonicity. For ease of exposition, assume that these assumptions are unconditionally satisfied for the binary IV, Z. Section 2.3 below discusses

[10] It is possible that the underlying decision-makers value factors that are not captured by the researcher-chosen $Y_i(D_i)$ (Kleinberg et al., 2018). If the relationship between such omitted factors and our chosen $Y_i(D_i)$ varies by group, this definition would not quantify differential treatment conditional on all unobservable factors. However, such gaps can still be interpreted as unwarranted disparities if conditioning on the chosen $Y_i(D_i)$ maps to a well-defined notion of fairness.

implementing our approach with common quasi-experimental designs that generate conditionally-random binary IVs, such as regression discontinuity (RD) and difference-in-difference (DiD) designs.

Definition 2 describes a time period-specific version of Definition 1, where the periods are delineated by the values of Z: $\triangle_{zy}$ is a group treatment gap that is specific to a given period and value of the treated outcome. Each such gap is composed of treatment rates that are conditional on time period, group and potential outcome (π_{zry}).

Definition 2. Differential treatment within a given period, conditional on $Y_i(1)$

$$\begin{aligned}\triangle_{zy} &= (E[D_i|Z=z, R_i=r_1, Y_i(1)=y] - E[D_i|Z=z, R_i=r_2, Y_i(1)=y]) \\ &= (\pi_{zr_1y} - \pi_{zr_2y}) \\ \triangle_z &= \sum_{y \in supp(Y_i(1))} Pr(Y_i(1)=y)\triangle_{zy}\end{aligned}$$

Our objects of interest are the period-specific estimates of discrimination that are conditional on having the same outcome if treated ($\triangle_z$). These are averages of the period- and treated outcome-specific gaps, weighted by the population prevalence of each value of the treated potential outcome ($Pr(Y_i(1)=y)$). These period-specific discrimination objects can also be differenced to measure changes in discrimination **due to the intervention**, $\triangle_{z=1} - \triangle_{z=0}$.[11]

To understand how to estimate our main object of interest, $\triangle_z$, note that its building blocks are treatment rates that are conditional on time period, group, and treated outcome (π_{zry}). As discussed earlier, since the treated outcome is selectively observed, we cannot directly condition on it to compute each π_{zry}. However, following Arnold, Dobbie, and Hull (2022), we re-write π_{zry} in Equation 1 using: 1) the definition of conditional expectations, and 2) the IV assumptions. The second line follows from the definition of conditional expectations, while the third line follows from the fact that $Y_i(1) \perp Z$.

$$\begin{aligned}\pi_{zry} &\equiv E[D_i|Z=z, R_i=r, Y_i(1)=y] \\ &= \frac{E[Y_i(1)=y|Z=z, R_i=r, D_i=1] \times E[D_i|Z=z, R_i=r]}{E[Y_i(1)=y|Z=z, R_i=r]} \\ &= \frac{\overbrace{E[Y_i(1)=y|Z=z, R_i=r, D_i=1] \times E[D_i|Z=z, R_i=r]}^{\text{Observed in data}}}{\underbrace{E[Y_i(1)=y|R_i=r]}_{\text{Unobserved}}}\end{aligned} \tag{1}$$

Equation 1 highlights how to quantify the time period-, group-, and potential outcome specific treatment rates (π_{zry}) used to estimate discrimination conditional on having the same outcome if treated. We need the following objects:

[11] Even though Z represents quasi-experimental variation, the cross-group gap in the impact of Z on D will not generally recover $\triangle_{z=1} - \triangle_{z=0}$ unless 1) potential outcomes are similar across groups or 2) the impact of Z on D is uncorrelated with potential outcomes. Such cross-group comparisons can also suffer from the pitfalls of conducting marginal outcome tests with discrete instruments (Canay, Mogstad, and Mountjoy, 2024).

1. $E[D_i|Z = z, R_i = r]$:
   Share of individuals of each group r treated in each period z

2. $E[Y_i(1) = y|Z = z, R_i = r, D_i = 1]$:
   Share of treated individuals with treated outcome y, for each group r and period z

3. $E[Y_i(1) = y|R_i = r]$:
   Prevalence of treated potential outcome y in each group's population

Objects 1 and 2 in the list above are observed in data—we see the share of individuals of each group who are treated and the outcomes realized for treated individuals. Object 3, the share of individuals of group r who would experience a given value of the treated potential outcome, would only be observed in a counterfactual where everyone of that group was treated. This share is an especially crucial element because it also provides a test of whether the distribution of potential outcomes varies by group. If the distributions differ by race, we cannot interpret the raw observed group differences in treatment as discrimination.[12]

Object 3 is typically not directly observable in the data unless unique institutional features generate a subsample where everyone is treated (and there is no discrimination in that sample by definition). In certain contexts, random assignment to supremely lenient decision-makers who treat almost everyone permit statistical extrapolations of $E[Y_i(1) = y|R_i = r]$ (Arnold, Dobbie, and Hull, 2022; Baron et al., 2023). However, there are many settings where measuring discrimination is of interest, but the empirical conditions needed to apply these existing approaches are not satisfied. For example, current extrapolation techniques are difficult to apply if decision-makers are not randomly assigned or they are unobserved in the data, both of which might be common in empirical work. In our main empirical application, criminal misdemeanor defendants in King County are not randomly assigned to prosecutors. Thus, using existing approaches in such a setting to assess if the distribution of potential outcomes differs by group and measure discrimination would be difficult.

Below, we provide an approach to measuring discrimination that greatly expands the settings where researchers can measure discrimination conditional on potential outcomes. We incorporate insights from the IV and marginal treatment effects literatures to estimate $E[Y_i(1) = y|R_i = r]$, which we need to understand if the groups to be compared differ in terms of the potential outcomes if treated. We then measure discrimination accounting for any differences in these distributions following Equation 1 and Definition 2.

2.2 Implementation with a binary instrument

We describe our framework in the context of our empirical application studying racial discrimination in misdemeanor prosecution in Washington State. Individuals belong to either 'White' or 'Minority'

[12] We also use Object 3, along with each group's shares of the population (p_r), to compute the underlying prevalence of treated potential outcome y in the population: $E[Y_i(1) = y] = p_{r_1} E[Y_i(1) = y|R_i = r_1] + p_{r_2} E[Y_i(1) = y|R_i = r_2]$, which we use to aggregate each $\triangle_{zy}$ to estimate discrimination in a given period.

groups, denoted by $R_i \in \{w, m\}$, and the treatment decision is prosecution.[13] We continue our discussion of estimating racial differences in prosecution rates for individuals who would have the same outcome if treated, i.e., prosecuted (see Appendix B.1 for extensions to measure discrimination conditional on the untreated potential outcome, treatment effect, or two potential outcomes, e.g., recidivism risk and case quality).

Individuals, indexed by i, are chosen for treatment, D_i. In the context of prosecution, this treatment decision is made by multiple agents who influence case outcomes, including prosecuting attorneys and judges, rather than by the individual i. Let $D_i = 1$ if an individual's case is prosecuted, and $D_i = 0$ if an individual's case is dismissed. Let the potential outcomes be binary indicators for whether an individual commits a new offence in the future, after prosecution or dismissal, i.e., $Y_i(D_i) \in \{0, 1\}$. While we use binary potential outcomes in this section for simplicity, the expressions in Definition 2 and Equation 1 accommodate multi-valued potential outcomes. Finally, let Z be a binary instrument that satisfies the standard IV assumptions listed in Section 2.1 and shifts the rate of prosecution. In the context of our application in Section 3, let Z represent periods before and after an unanticipated budget cut that sharply **reduced** prosecution rates.

Recall that we need to estimate each time period-, group-, and potential outcome-specific treatment rate (π_{zry}) in Equation 1 to quantify the discrimination estimands in Definition 2. The key challenge is that the denominator, $E[Y_i(1) = y|R_i = r]$, is unobserved. Overcoming this challenge requires estimating the proportion of each group r that would realize treated outcome $Y_i(1) = y$ if everyone in that group were treated. Since $Y_i(1)$ is binary, this proportion coincides with the average outcome that would be realized if everyone in group r were treated. We next show how to use the binary instrument Z to estimate bounds and point estimates of $E[Y_i(1) = 1|R_i = r]$.

Under the standard IV assumptions discussed earlier, the variation from the binary instrument Z partitions the population into three "compliance groups" (always takers, compliers, and never takers) and identifies their population shares as well as certain average potential outcomes (Angrist, Imbens, and Rubin, 1996). If Z shifts treatment for both racial groups, these quantities are identified separately by race group. For each race, we directly observe the proportions of always takers (p_A) and compliers (p_C) in the data by examining the share of the population that would receive treatment regardless of the reform and would only receive treatment because of the reform, respectively. In the example here, since the reform decreases prosecution rates (the treatment), p_A is the share of people prosecuted after the budget reform and p_C is the change in prosecution rates due to the reform. Since always takers, compliers, and never takers partition the population of group r, the share of never takers is $p_N = 1 - p_A - p_C$.

The variation from the binary IV also provides estimates of average potential outcomes for a subset of the "compliance groups". Focusing on treated potential outcomes again, first note that in the prosecution application, the average treated outcome for always takers is the average outcome of people treated (i.e., prosecuted) after the budget cut. Second, note that the group

[13] This definition of racial groups, rather than a White–Black comparison, is motivated by the context of Washington State, which we describe in detail in Section 3.

of people treated before the budget cut consists only of compliers and always takers. Hence, the average outcome of people treated before the reform is a weighted average of treated outcomes for compliers and always takers. Using these two averages, along with the population shares of always takers and compliers, we estimate the average treated outcomes of compliers (Imbens and Rubin, 1997). This information recovers average treated outcomes for two of the three "compliance groups" that partition the population of a given racial group: always takers and compliers. However, we do not observe the average treated outcomes of never takers, since they are never treated by definition. That is the final piece to estimate the average outcome that would be realized if everyone of each racial group were treated, $E[Y_i(1) = 1|R_i = r]$.

We estimate bounds (or point estimates) for the average treated outcomes of never takers by placing restrictions on the relationship between treatment propensity and average treated outcomes.[14] Each "compliance group" is defined by its propensity to be treated: always takers are more likely to be treated than compliers, who are in turn more likely to be treated than never takers. Figure 1 depicts a hypothetical example where always takers and compliers are roughly 70% and 20% of the population respectively. In this example, compliers have greater treated outcomes than always takers. In the context of prosecution, this implies that on average, marginally prosecuted individuals are more likely to commit a new offence if prosecuted than those who are very likely to be prosecuted.[15] We use this estimated relationship to infer the treated outcomes of never takers.

In Panel a) we assume that **average** treated outcomes are weakly monotonic in the treatment propensity of "compliance groups", following Mogstad, Santos, and Torgovitsky (2018) and Kowalski (2023a). This assumption extends the relationship between always takers' and compliers' average treated outcomes to never takers' average treated outcomes. In this example, the assumption implies that the average treated outcomes for never takers must be weakly greater than the average treated outcomes for compliers, pinning down the lower bound for never takers' average outcomes. That is, since $E[Y_i(1)|\text{ A}, R_i = r] < E[Y_i(1)|\text{ C}, R_i = r]$, weak monotonicity implies that $E[Y_i(1)|\text{ C}, R_i = r] \leq E[Y_i(1)|\text{ N}, R_i = r]$, where A, C, and N denote always takers, compliers, and never takers. Next, since $Y_i(1) \in \{0, 1\}$, we have that $E[Y_i(1)|\text{ N}, R_i = r] \in [E[Y_i(1)|\text{ C}, R_i = r], 1]$. Similar arguments hold for when the relationship between the average treated outcomes for always takers and compliers is reversed, i.e., if $E[Y_i(1)|\text{ A}, R_i = r] > E[Y_i(1)|\text{ C}, R_i = r]$.

Finally, we compute a weighted average of 1) the bounds on the average treated outcomes for never takers and 2) the point estimates for the average treated outcomes for compliers and always takers, where the weights are the shares of each compliance group. Together, these yield bounds on the average treated outcomes for each racial group, $E[Y_i(1) = 1|R_i = r]$.[16] The formal description

[14] When both a natural experiment and random assignment to decision-makers are present, the approach that identifies outcomes using information for a larger proportion of treated (or untreated, if conditioning on the untreated potential outcome) individuals will require less extrapolation and involve less extrapolation error.

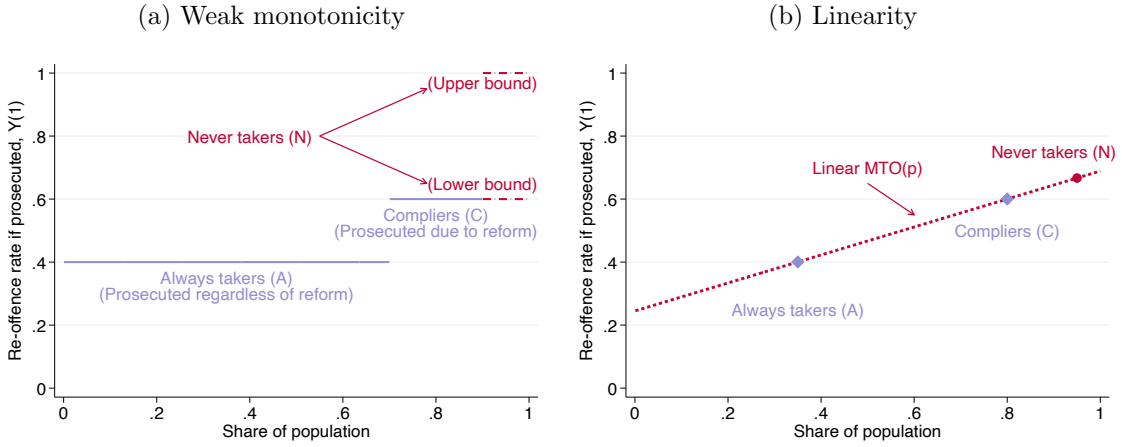
[15] Such a pattern might arise if prosecution was focused on individuals who would reduce their future offending if prosecuted.

[16] This bounding approach also applies if potential outcomes are multi-valued. In such a case, we require an estimate of the prevalence of each possible value of the potential outcome, $E[Y_i(1) = y|R_i = r]$, to identify the discrimination estimand in Definition 2. The logic underlying the bounding approach will still hold, since each expectation $E[Y_i(1) = y|R_i = r]$ represents a proportion and is hence also bounded between $[0, 1]$.

of the bounds for $E[Y_i(1) = y|R_i = r]$, described in Equation 2 for the running example here, highlights the conditions under which these bounds (and consequently the bounds on discrimination conditional on the treated potential outcome) are narrower. These bounds are narrower when never takers are less common (reducing the share of the population whose treated outcomes we need to bound) and when compliers' average treated outcomes are different from those of always takers (reducing the vertical width of the bounds for never takers' average treated outcomes).[17]

$$\begin{aligned} &\text{Lower bound: } p_A E[Y_i(1)|\text{ A}, R_i = r] + (p_C + p_N)E[Y_i(1)|\text{ C}, R_i = r] \\ &\text{Upper bound: } p_A E[Y_i(1)|\text{ A}, R_i = r] + p_C E[Y_i(1)|\text{ C}, R_i = r] + p_N \end{aligned} \tag{2}$$

Figure 1: Identifying the average treated outcome with a binary IV



Note: This figure uses simulated data. Lower values of the x-axis denote individuals who are more likely to be treated. $Y(1)$ denotes the treated potential outcome. The diamonds and dots in Panel b) reflect outcomes of the median individual in that group. Solid lines and diamonds represent moments observed in the data, and dashed lines and circles represent objects that are extrapolated.

Our approach to bounding average potential outcomes is similar to the "performance bounding" approach in Jordan (2024), who studies racial discrimination in felony review in a context where prosecutors are randomly assigned cases. There, the average outcome for never takers is bounded by the average outcome for always takers and compliers together, which is motivated by modeling the underlying objectives of prosecutors. The bounds in our case are provided by the average outcome for compliers and represent restrictions on the aggregate behavior induced by the policy reform. These approaches will meaningfully diverge in practice when average outcomes for compliers and always takers are quite different from each other, as is the case in our empirical application.

It is important to note that the restrictions we place to bound average treated outcomes do not require assuming the underlying decision-makers focus on a single narrow objective. Rather, our approach lets decision-makers adopt a range of multi-dimensional models as long as their decisions

[17] Analogously, the bounds for average untreated outcomes are narrower when always takers are less common and when never takers' and compliers' untreated outcomes are very different.

generate patterns that imply that **average** potential outcomes for each of the "compliance groups" are weakly monotonic across the groups in the order of their treatment likelihood. In the context of prosecution, the extrapolation in Panel a) of Figure 1 assumes that never takers, who are least likely to be prosecuted, are at least as likely as compliers to re-offend if prosecuted. Thus, the assumption would be satisfied if prosecution was targeted towards people for whom prosecution reduces future crime. Weak monotonicity also lets prosecutors selectively dismiss cases they believe are most deserving of a second chance even if those people also have a high baseline risk of recidivism (e.g., youth).

On the other hand, this assumption might be violated if other inputs into prosecution decisions generate contradictory patterns. For example, say that prosecutors are also less likely to pursue defendants represented by private attorneys, such that all the never takers' cases have private representation. If individuals with private representation never re-offend if prosecuted, the weak monotonicity assumption could be violated. In such situations, an alternative approach is to bound never takers' treated outcomes between 0 and 1—the widest logically possible bounds (Manski, 1989).

Instead of bounds, we can also obtain point estimates of discrimination if we restrict the relationship between the underlying treatment propensity and treated outcomes to be linear. Panel b) of Figure 1 demonstrates this, where the diamonds plot the treated outcomes for the median always taker and complier against their respective treatment propensities.[18] Assuming this relationship is linear lets us extrapolate the treated outcomes across the support of the treatment propensity and point identify the treated outcomes of never takers. This restriction in turn identifies the marginal treated outcome function which we would integrate to estimate $E[Y_i(1) = 1|R_i = r]$ (Heckman and Vytlacil, 2000; Brinch, Mogstad, and Wiswall, 2017; Mogstad, Santos, and Torgovitsky, 2018; Kowalski, 2023b).[19] Since this approach involves assuming that the relevant marginal treatment response function is linear, it places stronger restrictions on behavior than the previous partial identification approach. We illustrate this point-identification approach in the context of racial discrimination in incarceration decisions in Appendix B.2, which sketches a simple model of selection, describes the required assumptions, and implements a brief empirical example.

Implementing this procedure separately by race group r provides either bounds or point estimates for the average outcome that would be realized if everyone of each group were treated, $E[Y_i(1) = 1|R_i = r]$. As mentioned earlier, this is the final object that we need to estimate discrimination that is conditional on treated potential outcomes, which is reproduced for the case of binary treated outcomes in Equation 3.

[18] Linearity and the uniformity of the underlying latent index determining treatment implies that the median outcome of each compliance group has the average treated outcome of that compliance group (Kowalski, 2023b).

[19] Linearity assumptions can identify all the marginal treatment response functions. We would identify the marginal untreated outcome function by using the average untreated outcomes for compliers & never takers to extrapolate the untreated outcome for always takers. Along with the marginal treated outcome function, this identifies the marginal treatment effect function.

$$
\begin{aligned}
\pi_{zr1} &= \frac{\overbrace{E[Y_i|Z=z,R_i=r,D_i=1]\times E[D_i|Z=z,R_i=r]}^{\text{Observed in data}}}{\underbrace{E[Y_i(1)=1|R_i=r]}_{\text{Extrapolated}}} \\
\pi_{zr0} &= \frac{\overbrace{E[(1-Y_i)|Z=z,R_i=r,D_i=1]\times E[D_i|Z=z,R_i=r]}^{\text{Observed in data}}}{\underbrace{1-E[Y_i(1)=1|R_i=r]}_{\text{Extrapolated}}} \\
\triangle_{zy} &= \pi_{zwy} - \pi_{zby} \\
\triangle_z &= \sum_{y\in\{0,1\}} Pr(Y_i(1)=y)\triangle_{zy}
\end{aligned}
\tag{3}
$$

We observe the re-offence rate among prosecuted individuals in each period and race in the data: $E[Y_i|Z=z,R_i=r,D_i=1]$. We also observe the prosecution rate for each period and race: $E[D_i|Z=z,R_i=r]$. Plugging the bounds/point estimates for $E[Y_i(1)=1|R_i=r]$ into the first two lines of Equation 3 generates bounds/point estimates for each period-, race-, and potential outcome-specific treatment rate, π_{zry}. Using π_{zry} in the third line yields bounds/point estimates for the period- and potential outcome-specific discrimination $\triangle_{zy}$. We then construct the period-specific discrimination estimates, $\triangle_z$, as a weighted average of the period- and potential outcome-specific discrimination, where the weights are defined by the prevalence of the treated outcome in the population, $Pr(Y_i(1)=1)$ (fourth line of Equation 3).

Appendix B.1 extends this approach to measure discrimination while simultaneously conditioning on treated and untreated outcomes (which also lets us condition on treatment effects) or two distinct potential outcomes. To do so, we need to identify versions of the treatment rates in Equation 1 that are conditional on two values of potential outcomes—e.g., prosecution rates for people who would re-offend if prosecuted ($Y_i(1)=1$) but not if dismissed ($Y_i(0)=0$). We identify such conditional treatment rates in two steps. First, we use the approach described above to extrapolate population averages for each potential outcome of interest. Second, we bound the covariance between the potential outcomes using information from the extrapolations and observed means, following Hössjer and Sjölander (2022). This procedure bounds moments of the joint distribution, which we use to quantify treatment rates that are conditional on two potential outcomes and then measure discrimination.

2.3 Implementation with common quasi-experimental variation

So far, we have discussed using an unconditionally random binary IV to assess if treated outcomes differ by group (e.g., race) and then estimate discrimination, conditional on treated outcomes. Here, we discuss the practical details of implementing our approach with common quasi-experimental designs that generate conditionally random binary IVs, e.g., regression discontinuity

(RD) or difference-in-difference (DiD).

Regression discontinuity

As is well known, RDs are well-suited to identify treatment effects and average potential outcomes at a particular cut-off. The simplest way to implement our approach with an RD is in the spirit of the local randomization approach, which typically involves only using information in a small region just below and just above a cut-off (Cattaneo, Idrobo, and Titiunik, 2024). In such a setup, the binary instrument will be an unconditionally random indicator for being above or below the cut-off, letting us follow our approach as discussed above. However, the sample size reduction from limiting to a very small window around a cut-off will typically result in power issues.

When using our approach with an RD, we can use data from a larger window around the cut-off with some minor adjustments. To see how, consider the following expression of an RD design: $X_i = \alpha + \beta\,\mathbb{I}(S_i{>}c) + \delta_1 f(S_i) + \delta_2\,\mathbb{I}(S_i{>}c) \times g(S_i) + \varepsilon_i$. Here S represents the running variable, $f(\cdot)$ and $g(\cdot)$ govern the relationship between S and the outcome, c represents the cut-off, and the binary instrument Z is given by $\mathbb{I}(S_i{>}c)$. When estimating this equation with X_i as an indicator for treatment, α estimates the share of individuals at the cut-off receiving treatment despite having $Z = 0$—in other words, the share of always takers. β represents the share of individuals at the cut-off receiving treatment due to the instrument, pinning down the share of compliers. Next, consider estimating a version of the expression above on the set of treated individuals, this time with X_i representing the outcome of interest. In this case, α pins down the average treated outcome for always takers at the cut-off. $\alpha + \beta$ pins down a weighted average of treated outcomes for always takers and compliers at the cut-off.

As in the case with a simple binary IV, we have all the information we need to 1) back out the average treated outcomes for compliers, 2) obtain bounds or point estimates for the average treated outcomes in the population, and 3) use that to measure discrimination at the RD cut-off conditional on potential outcomes. Appendix B.3 illustrates using an RD design with our approach, through an empirical example studying socio-economic discrimination in the decision to allow public school students to progress to the next grade.

While this implementation uses observations away from the cut-off, the resulting estimates of discrimination conditional on potential outcomes are still informative only for observations at the RD cut-off rather than away from it. Extending the approach to measure discrimination in a small window around the cut-off requires two additional assumptions (described in detail in Appendix B.4). First, the running variable alone cannot influence treatment. Second, the average outcome if everyone at the cut-off were treated would need to be the same as the average outcome if everyone above or below the cut-off were treated.[20] Since the running variable often has a tight relationship with both treatment and potential outcomes in many settings, these assumptions are

[20]These assumptions are stronger than those in typical approaches to extrapolate away from RD cut-offs (e.g., Angrist and Rokkanen, 2015; Cattaneo et al., 2021; Ricks, 2022) because our focus is on extrapolating average potential outcomes rather than treatment effects.

generally strong and may be less likely to be satisfied.

Difference-in-difference

Consider a setting, as in our misdemeanor prosecution application, where the key variation comes from the timing of a budget reform adopted in one county but not others. Mapping this variation to the IV framework that is the core of our approach, individuals rather than counties select into treatment (i.e., prosecution) and time might independently influence potential outcomes and treatment. With this time variation, IV designs in the spirit of two-way fixed effects that condition on indicators for county and time period will not identify the proportions and average potential outcomes for each of the "compliance groups", information that we need to measure discrimination.

Under additional assumptions, we can account for the time trends in the county that *adopted the reform*, which we do in our application studying discrimination in misdemeanor prosecution. Focusing on the treated potential outcome and the bounding approach, we account for time trends by using the change in treated outcomes in counties that *did not adopt a reform* as an estimate of the change in treated outcomes the county that *adopted the reform* would have experienced if it had not adopted the reform. This adjustment assumes that for individuals in each racial group: 1) time alone does not influence treatment status, and 2) time trends in average treated outcomes are the same for always takers and compliers, and are independent of county (in a window around policy adoption).[21,22,23] Appendix B.5 formally discusses the assumptions and adjustment, and also discusses extensions to condition on untreated potential outcomes. This adjustment essentially provides a way to map DiD designs to existing IV methods of estimating average potential outcomes for each "compliance group" (Imbens and Rubin, 1997).

After correcting for the effects of time in the county that adopted the reform, we only use information from that county and define the instrument as we would in the simple binary case, where Z represents periods before and after the reform. Note that since there are time trends in potential outcomes, the average treated outcomes for each compliance group and the average outcomes if everyone were treated will vary by time period.

The logic of this approach also extends to DiD with staggered policy adoption. Following prior work on estimating policy impacts in settings with staggered adoption, we can pair each county that adopts a policy ('adopter') with a set of counties that never adopted or have not yet adopted the policy ('not-yet-adopter') (Cengiz et al., 2019). Then, the second assumption described above needs to hold between each 'adopter' and its associated set of 'not-yet-adopters'. Combining the IV

[21]The second assumption is stronger when using our approach to obtain point estimates of discrimination, which requires assuming that the outcomes of always takers, compliers, and never takers are linearly related. Hence, the assumption of equal of time trends across always takers and compliers must extend to never takers as well.

[22]While this second assumption is in the spirit of parallel trends, it contrasts with the usual DiD implementations in which everyone in counties where a policy occurs are considered 'treated' by the policy, while the rest are considered 'untreated'. There, the standard parallel trends assumption identifies impacts of the policy by assuming parallel trends in the average untreated potential outcomes between the two counties.

[23]These assumptions are similar to those underlying the "time-corrected" Wald estimand in De Chaisemartin and D'Haultfœuille (2018), which allow identification of the local average treatment effect but not of the average potential outcomes of each "compliance group" separately.

extrapolation approach described earlier with these DiD adjustments for each 'adopter' provides estimates of discrimination for each 'adopter'. These estimates can be aggregated across 'adopters' to construct an average measure of discrimination across all 'adopters'.

2.4 Recap: Step-by-step guide

In this section we have discussed how to use a natural experiment to estimate discrimination between two groups, accounting for any underlying differences in the distribution of potential outcomes.[24] Implementing the approach requires the following:

1. A natural experiment that generates a binary IV
2. A potential outcome, $Y_i(D_i)$, that corresponds to a notion of fairness. That is, the choice of $Y_i(D_i)$ should be such that group differences in treatment between people with the same $Y_i(D_i)$ can be interpreted as an unwarranted disparity.
3. Use the natural experiment to estimate whether the underlying distributions of potential outcomes differ between the groups. For example, if $Y_i(D_i)$ is binary and we want to condition on the treated potential outcome, this involves comparing the average outcomes if everyone of each group were treated.
    - To obtain bounds for this object, either: i) assume average potential outcomes are weakly monotonic across always takers, compliers, and never takers or ii) extrapolate the unobserved outcomes using the largest and smallest logically-possible values for the outcome.
    - Obtain point estimates by assuming potential outcomes are linear in the underlying likelihood of treatment
4. Estimate average discrimination conditional on potential outcomes following Equation 1.
    - Even if the average potential outcomes from Step 3 are not significantly different across groups, estimates from Step 4 may still provide useful information about discrimination given different possible values for the average potential outcomes, especially in the case of partial identification.

In the next section we apply this approach to study racial discrimination in prosecution. We adopt the partial identification approach since the data are inconsistent with the linearity restriction required for point identification. We first bound the average potential outcomes separately

[24] So far we have discussed estimating aggregate measures of discrimination. One can also use our approach to quantify decision-maker-specific discrimination estimates with large enough samples and first stages within the subsample for each decision-maker, j. Repeating our approach within each j's subsample yields estimates of the average potential outcomes in each j's subsample, $E[Y_i(D_i)|j]$. Decision-makers with similar $E[Y_i(D_i)|j]$ observe subsamples with similar potential outcomes. j-specific discrimination estimates are comparable, in a 'selection-on-unobservables' strategy, among such a subset of decision-makers.

by race, and then plug these bounds directly into the expression for the average period-specific discrimination ($\triangle_z$).[25] We compute the bounds for $\triangle_z$ by searching over a grid defined by the combinations of the average potential outcome bounds for each racial group.

For inference, we generate 95% bootstrapped confidence intervals for these bounds using a Bayesian bootstrapping procedure (Rubin, 1981). We use weights randomly drawn from $\Gamma(1,1)$ to compute all the moments in the estimation procedure, enforcing the weak monotonicity restriction within each re-weighted bootstrap sample. Finally, we report confidence intervals for the true underlying parameter (rather than each bound) using the resulting bootstrap distribution (Imbens and Manski, 2004).

3 Racial discrimination in misdemeanor prosecution

In this section we use quasi-experimental variation from a cut to the King County prosecutors' budget in Washington State to measure racial discrimination in the decision to prosecute individuals arrested for misdemeanors. Misdemeanors represent the most common form of contact with the criminal court system (Stevenson and Mayson, 2018) and the decision to prosecute cases like these can have long-lasting adverse impacts on individuals' lives (Leasure, 2019; Mueller-Smith and Schnepel, 2021; Agan, Doleac, and Harvey, 2023).

We first find that in this setting, potential re-offence outcomes if prosecuted differ by race. Accounting for these racial differences in potential outcomes, we find no evidence of discrimination in prosecution before the budget cut. After the budget cut, which sharply reduced prosecution rates, we find that white defendants were more likely to be prosecuted than minority defendants. We find suggestive evidence that this gap is driven by prosecutors dropping cases that were likely low quality, which were more prevalent among minority defendants. Crucially, misdemeanor cases are typically not randomly assigned to prosecutors in King County, ruling out existing random assignment-based tools to measure discrimination here.[26]

3.1 Natural experiment: King County budget reform

We study racial discrimination in misdemeanor prosecution in King County, Washington (Seattle metropolitan and suburban areas) using administrative records on all criminal cases from 2002–

[25] Note that one could also bound average period-specific discrimination by 1) first constructing gaps using the bounds on the group- and potential outcome-specific treatment rates, 2) computing the average, and then 3) taking the minimum and maximum. This will differ from our approach of plugging in the bounds on average potential outcomes directly into the equation for average period-specific discrimination because minimum/maximum are not linear functions. We prefer directly plugging bounds on average potential outcomes into the expression for $\triangle_z$ in Equation 3 since our main goal is estimating discrimination conditional on potential outcomes.

[26] Correspondence with the King County Prosecuting Attorney's Office suggests most cases submitted by police are assigned to an attorney (or a team of attorneys) based on offence type, experience, and workload, who decide whether to a file the case. The filing decision is then approved by supervising attorneys. While some subsequent proceedings for filed cases may be quasi-randomly assigned (e.g., arraignment based on a stipulated calendar), prosecutors for many steps are often assigned at the discretion of supervising attorneys.

2014 from the Washington Administrative Office of the Courts.[27] We consider an individual as having their case prosecuted if their case **did not** meet the following condition: dismissed without requiring any punishments.[28] Our primary definition of punishment excludes fines, but we assess robustness to including them.

We focus on differences in prosecution between white (non-Hispanic) and 'minority' defendants. This comparison is motivated by the fact that the population that has contact with the Washington criminal legal system is quite diverse. A large proportion of non-white defendants are of Native Hawaiian and Pacific Islander descent—these groups often face disadvantage of various forms and are over-represented in the criminal legal system in Washington State and the Western United States in general (Hu and Esthappan, 2017; Buch and Borkholder, 2020; Malott, 2024). Nevertheless, we later demonstrate that our results are robust to comparing prosecution rates between white (non-Hispanic) and Black & Hispanic defendants.

As discussed in Section 2.2, we need a natural experiment that shifts prosecution rates to assess if unobserved potential outcomes vary by race, and then estimate discrimination conditional on the unobserved potential outcomes. We use the fact that King County, facing a $60 million budget shortfall in September 2010, cut the Prosecutors' Office budget by approximately $3.9 million, the equivalent of 33 full-time employees (Constantine, 2010). In response, the Prosecutors' Office warned that the unanticipated reduction in resources would reduce their ability to prosecute challenging and time-consuming cases, and that they would have to focus resources on high-priority offences (Ervin, 2010). The County tried to mitigate the budget cuts' impact on the criminal legal system by holding a referendum to raise funding via a sales tax increase. However, the referendum failed in November 2010, consigning the King County Prosecutors' Office to the new budget realities (Ballotpedia, 2010).[29]

We use the budget reform with our discrimination estimation approach to study racial discrimination in prosecution. The sharp contraction to prosecutorial resources should result in many cases being dropped, especially given the Prosecutors' Office's public statements. A shift in prosecution rates would let us partition each racial group into always takers, compliers, and never takers for prosecution. We would then examine how the average prosecuted outcomes vary across racial groups, and then account for any racial differences in average prosecuted outcomes.

We isolate the quasi-experimental variation using a difference-in-difference strategy that com-

[27] Since these data consist of individuals in the court records our estimates are representative of individuals who have been arrested and whose cases have been sent by police to the prosecutors' office.

[28] Our data do not allow us to accurately distinguish between situations where prosecution was pursued and: individuals were convicted, prosecution failed, and charges were dismissed upon successful completion of a sentence. Our definition considers all of the above scenarios as 'prosecution'.

[29] We rule out that the change in the Seattle City Attorney, who pledged to reduce racial disparities and prosecution of minor offences, on January 1, 2010 poses a confounding threat. The City Attorney has jurisdiction over Seattle's local municipal courts (and county prosecutors do not), which we exclude from our sample of case dispositions. Hence, direct impacts of the City Attorney change are unlikely to be present in our analysis. Since this change occurred before the county budget cut, any indirect effects of the City Attorney's reforms on the broader courts in our sample should show up as differential pre-trends in the year leading up to the budget reform. We do not find evidence of such differential pre-trends during this period in our event studies examining prosecution, recidivism, or caseload composition (discussed below).

pares changes in prosecution and recidivism rates around the budget reform in King County, relative to changes in the adjacent counties unaffected by the reform (Chelan, Kitsap, Kittitas, Pierce, and Snohomish). We construct our analysis sample using criminal cases disposed in the District Courts of these counties. County prosecutors who work in District Courts typically hear criminal misdemeanor cases of varying severity. Given the messaging from the Prosecutors' Office regarding the types of cases they will find difficult to pursue, cases in District Courts are most likely to be affected by the prosecutorial budget cuts. We limit to misdemeanor cases disposed in the relevant District Courts between October 2008 and September 2012, a two-year span on either side of the budget cut announcement.[30]

We construct re-offence outcomes and measure criminal history using information on cases filed by law enforcement in Washington State, including felony offences and offences filed outside of the October 2008 and September 2012 interval. We measure re-offending by tracking whether a defendant appears again in these data after disposition. This measure of re-offending will not capture undetected criminal activity or arrests that police did not forward to prosecutors. However, it will capture new arrests that police forward to prosecutors, including cases that prosecutors choose not to pursue.

The final sample, described in Table 1, consists of around 120,000 unique cases. 30% of the sample consists of non-white defendants, who broadly represent the diverse population of Washington State—43% are Black, 32% are Hispanic and 18% are Asian American or Pacific Islander (AAPI). We refer to the group of non-white defendants as 'minority' defendants. Women make up almost a quarter of the sample, and the sample consists of a wide range of individuals in terms of age and criminal background. The average defendant is almost 35 years old and 47% of individuals have had at least one conviction in the past (averaging 3.5 prior convictions conditional on having any). Prosecution rates are quite high (over 80%) and individuals are unlikely to face jail time if they prosecuted. Most sentences involve fines, probation, or alternative treatments (e.g., substance abuse treatment). Only 9% of individuals prosecuted in King County (and 6% overall) in our sample were sentenced to any jail time and this is higher for minority defendants than for white defendants (12 vs 7%). However, any racial differences in sentencing amounts to a negligible difference in the number of days spent in jail—among those who serve any time in jail, the average sentence is around 40 days long. Additionally, only 0.5% of prosecuted cases in our sample involve the maximum jail sentence allowable for misdemeanors by Washington statute (one year).

Estimating the first stage & shifts in recidivism outcomes

We first compare the changes in prosecution rates before and after the King County budget reform to changes in prosecution rates in adjacent counties that were unaffected by the budget reform. This involves estimating the specification in Equation 4, where D_{itg} denotes whether defendant i was prosecuted in quarter t and g denotes whether the case was disposed in King County or the adjacent counties. We investigate the changes in prosecution rates separately for

[30] If an individual has cases filed on multiple dates within this time frame, we only include the first case to ensure that the probability of multiple appearances in our sample is not a function of prosecution decisions.

Table 1: Characteristics of the Washington District Court sample

	Overall	King County	Adjacent Counties
N	120,516	50,503	70,013
Demographics			
White (Non-Hispanic)	0.72	0.65	0.78
Black	0.12	0.16	0.09
Hispanic	0.09	0.09	0.09
AAPI	0.05	0.08	0.03
Age at disposition	34.6	34.8	34.4
Male	0.74	0.73	0.74
Criminal history			
Any prior convictions	0.47	0.42	0.50
# prior | Any	3.8	3.5	3.9
Case outcomes			
Case prosecuted	0.86	0.82	0.88
White	0.86	0.83	0.87
Minority	0.85	0.81	0.89
Jail sentence | Prosecuted	0.06	0.09	0.05
White	0.06	0.07	0.05
Minority	0.08	0.12	0.04
Sentence length (Days) | Jail sentence	40.3	39.5	41.2

Note: Sample includes all criminal cases disposed in the District Courts in Chelan, King, Kitsap, Kittitas, Pierce, and Snohomish counties in Washington State between October 2008 and September 2012. For defendants with multiple dispositions in this time frame, we include only the first case. "AAPI" stands for Asian American or Pacific Islander.

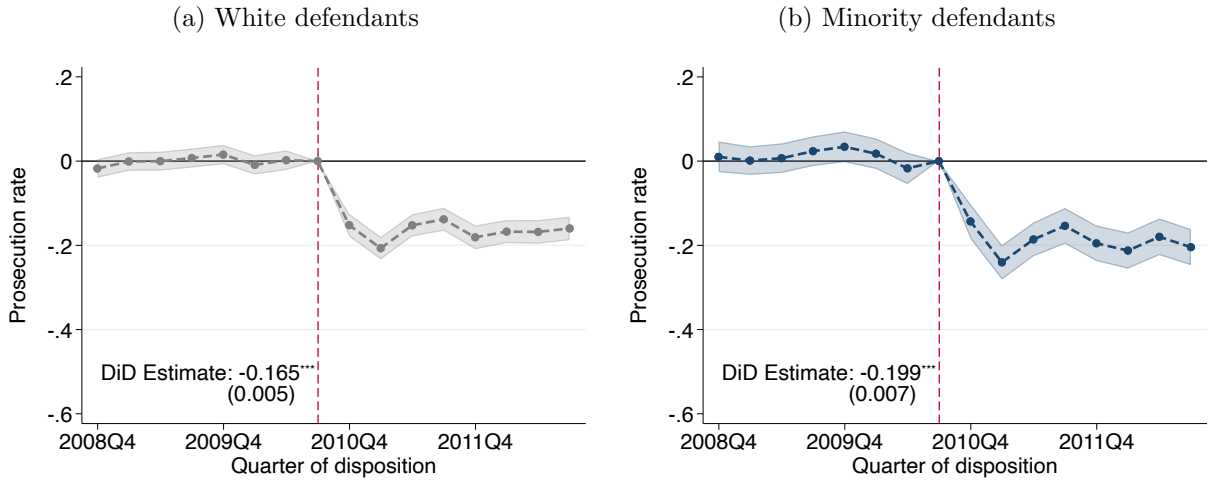
white and minority defendants to ensure that we have sufficient variation in prosecution rates in each racial subgroup.

$$D_{itg} = \theta_t + \delta \, \mathbb{I}[\text{King County}] + \sum_{k \neq -1} \beta_k \left(\mathbb{I}[t - \text{Budget Reform} = k] \times \mathbb{I}[\text{King County}] \right) + \varepsilon_{itg} \quad (4)$$

Figure 2 documents large drops in prosecution rates due to the King County budget reforms. Prosecution rates fall by 16.5 p.p. and 19.9 p.p. for white and minority defendants respectively (17.8% and 21.7% of pre-reform average prosecution rates in King County). Reassuringly, we do not see any evidence of differential pre-trends in prosecution rates, which provides initial evidence in support of a credible natural experiment (we discuss additional validity tests below). The magnitude of these shifts in prosecution rates pin down the complier share of white and minority defendants. Given the overall high rate of prosecution, these large shifts in prosecution suggest that we will have to bound outcomes for a relatively small share of the population.

We next examine whether these shifts in prosecution rates influenced individuals' re-offence outcomes, which will reduce the width of the bounds for never takers' prosecuted outcomes. Figure 3 repeats the event study approach, assessing the impact of the budget reform on the probability that a defendant re-offends one year after disposition. We see that not being prosecuted reduces one year re-offence rates by 3.1 p.p. for white defendants and 4.7 p.p. for minority defendants (13.1%

Figure 2: Impact of King County budget reform on prosecution rates



Note: Each Panel presents event study estimates investigating the impact of the King County budget reform. Sample includes all misdemeanor defendants as described in Table 1. 'DiD Estimate' pools the coefficients on relative time indicators and estimates $D_{igt} = \alpha + \delta_1 \mathbb{I}[\text{King County}] + \delta_2 Post_i + \beta^{DD} \mathbb{I}[\text{King County}] \times Post_i + \epsilon_{igt}$, where $Post_i = 1$ if the case is disposed on or after September 28, 2010, when the budget reform was announced. 95% confidence intervals are constructed using heteroscedasticity-robust standard errors.

and 15% of pre-reform average re-offence rates in King County).[31,32,33] Again, we see no evidence of pre-trends in re-offence rates. Figure A2 presents event study estimates showing how the reform impacts prosecuted outcomes—we later use the shift in prosecuted outcomes to estimate average prosecuted outcomes for always takers and compliers.

Assessing identifying assumptions

Next, we rule out various threats to the natural experiment's validity. A key concern is the presence of other concurrent policy or behavioral changes that could influence the determinants of crime and confound our estimates of the budget reform's effects. These might occur if other aspects of King County institutions, e.g., police or social services, were affected by the budget reform, or if prosecutors altered their behavior in ways other than prosecuting fewer cases. We assess how likely these concerns are through multiple exercises.

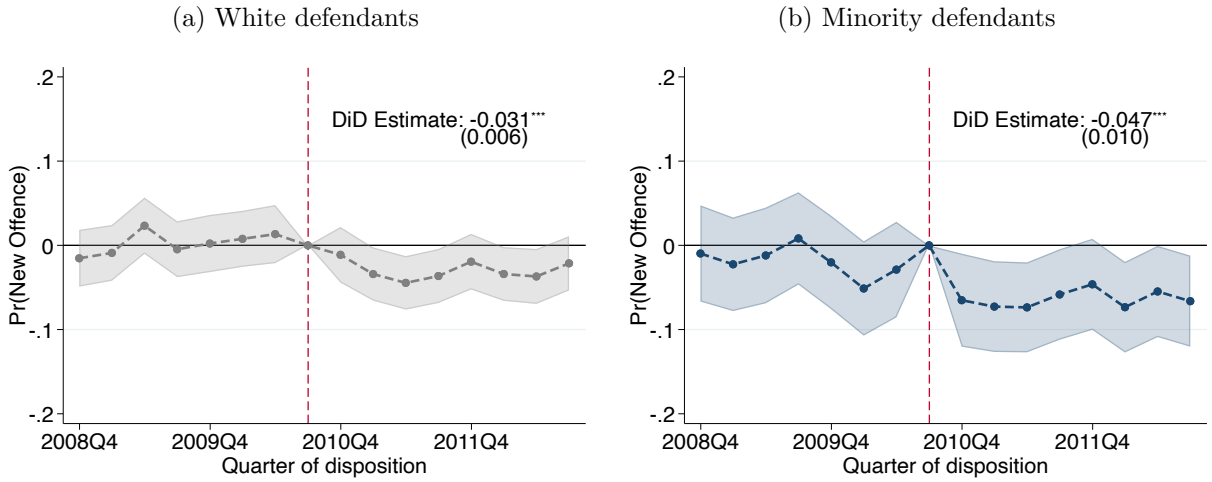
First, we test whether the composition of cases that prosecutors choose to accept from law

[31]These reductions are unlikely to be driven by incapacitation since only 9% of sentences for prosecuted defendants in King County involved jail. Among those with jail sentences, the average length was 40 days, much shorter than the one year horizon used to compute the outcome (see Table 1). As a result, potential outcomes are not mechanically censored by jail spells.

[32]These reduced form estimates are not driven by prosecutors systematically changing what cases they accept from law enforcement, e.g., by refusing all low priority cases. Figure A1 presents results excluding charges for offences commonly dropped right after the budget reform announcement (resisting arrest, criminal trespass, driving with a suspended licence, minor marijuana possession, reckless driving and DUI). Assuming these commonly-dropped charges are those that the prosecutor's office considers low-priority, we designate other charges as 'high-priority'. We see reductions in the probability of 'high-priority' re-offending that are similar to our baseline estimates in terms of their proportion of the relevant pre-reform means.

[33]The direction and magnitude of these estimates are consistent with recent work finding that avoiding prosecution for minor offences reduces future interactions with the criminal legal system (e.g., Mueller-Smith and Schnepel, 2021; Agan, Doleac, and Harvey, 2023).

Figure 3: Impact of King County budget reform on re-offence within one year



Note: Each Panel presents event study estimates investigating the impact of the King County budget reform. The outcome includes any new misdemeanor or felony offences committed anywhere in Washington State. Sample includes all misdemeanor defendants as described in Table 1. 'DiD Estimate' pools the coefficients on relative time indicators and estimates $Y_{igt} = \alpha + \delta_1 \mathbb{I}[\text{King County}] + \delta_2 Post_i + \beta^{DD} \mathbb{I}[\text{King County}] \times Post_i + \epsilon_{igt}$, where $Post_i = 1$ if the case is disposed on or after September 28, 2010, when the budget reform was announced. 95% confidence intervals are constructed using heteroscedasticity-robust standard errors.

enforcement changes due to the budget reform. We estimate a series of regressions where we compare the change in the share of individuals with a given baseline covariate before versus after the reform in King County to that same change in the adjacent counties. Figure A3 presents each of these coefficients and shows limited evidence of compositional shifts across multiple observable characteristics (Figure A4 presents the underlying event study patterns for each covariate). We see some evidence that individuals whose cases were accepted after the reform were 1.8 p.p. more likely to have been previously charged with an offence but this is a small shift relative to the pre-period mean in King County (3.5%).[34]

A related concern is whether law enforcement agents in King County were laid off or reduced their arrest effort (perhaps anticipating that prosecutors would stop prosecuting certain cases) due to the reform, and whether the reform affected other economic determinants of crime. In Figure A6 & Figure A7 we show that there are no significant reductions in either officer employment or different categories of arrests. Figure A8 similarly finds no differential changes in house prices, unemployment rates, or population, suggesting that the reform did not meaningfully impact other economic factors that might influence criminal behavior.

The results of these exercises suggest that the drop in prosecution rates and resulting reduction in recidivism are driven by the unanticipated budget reform that affected the operations of the

[34]This minor compositional shift might be due to the end of the federal investigation of Seattle Police Department (SPD) which resulted in lower SPD stops and arrests (Campbell, 2023). This compositional shift may arise if police focused on more serious offences. Consistent with this mechanism, Campbell (2023) finds the largest reductions in SPD activity after the investigation ended (December, 2011), which corresponds to when we see changes in defendants with a prior charge (see Panel c) of Figure A4). Our first stage and reduced form estimates are nearly identical if we exclude this time period (see Figure A5), and so we proceed with the full sample for our baseline analysis.

Prosecutors' Office, and not by concurrent policy factors or behavioral changes. Together with the stable pre-trends, these results provide strong evidence that the King County budget reform is a valid natural experiment. Next, we use this natural experiment to estimate racial differences in prosecution among individuals who would have the same re-offence outcome if prosecuted and then estimate the analog, conditioning on the re-offence outcome if dismissed.

3.2 Estimating discrimination in prosecution using the budget reform

As a refresher, we begin by mapping the objects in the empirical discussion to the potential outcomes framework discussed in Section 2.2. Individuals are considered 'treated' if they are prosecuted ($D_i = 1$) and considered 'untreated' if dismissed. Potential outcomes in each treatment state are binary indicators for whether an individual assigned to a given treatment state would re-offend within one year of disposition. We start by considering the re-offence outcome if prosecuted.[35] Z indicates periods around the King County budget reform, which reduced prosecution rates. As described in Section 2.2, the first step is to use the quasi-experimental variation from the budget reform to estimate the average re-offence outcome that would be realized if all defendants of each racial group were prosecuted. These objects would help us understand whether the unobserved potential outcomes vary by race, in which case we adjust the discrimination estimates for such differences.

Given the time-varying nature of potential outcomes in a DiD setting, we first need to adjust for any trends in potential outcomes due to the passage of time. Following the DiD adjustments discussed above, we use the trend in re-offence outcomes among prosecuted individuals in adjacent counties as an estimate of the time trend in outcomes that prosecuted individuals in King County would have experienced had the budget reform not occurred. As described in Section 2.3 and Appendix B.5, this adjustment is valid under two additional assumptions. In this context, we need the following conditions to hold within each race group:

A1 Without the reform, time does not shift individuals' prosecution status.

A2 Without the reform, re-offence outcomes if prosecuted, $Y_{it}(1)$, would trend similarly for always takers & compliers and is independent of county.

While these assumptions are fundamentally untestable, Appendix A.2 provides suggestive evidence that they are not violated in this setting. To examine **A1**, we use pre-period data and find negligible shifts in prosecution rates over time across covariate subgroup (gender, criminal history, age)-by-race cells (see Figures A25 and A26). These results build confidence that individuals are not shifting treatment status over time. To examine **A2**, which requires parallel trends to hold between

[35] Any racial differences in re-offence outcomes if prosecuted might be partially driven by racial differences in sentencing or future police interaction, complicating the interpretation of discrimination conditional on re-offence outcomes if prosecuted. To address this, we show that our eventual discrimination estimates are robust to i) comparing defendants with similar baseline likelihood of re-offending (by conditioning on the re-offence outcome if dismissed), ii) conditioning on re-offences for crimes that are less likely to involve police re-arrest discretion.

always takers and compliers, we test for differential pre-trends by covariate subgroup, since baseline characteristics might be correlated with being an always taker or complier. Using pre-period data, we find limited evidence of differential trends in $Y_{it}(1)$ across counties within the various subgroups (Figure A27) and within counties but across subgroups (Figure A28). While these patterns are not definitive evidence that these assumptions are satisfied, they build credibility that they are reasonable in this setting.

3.2.1 Baseline estimates of discrimination

We now use the variation from the budget reform DiD to estimate if there are racial differences in the average re-offence outcome that we would see if all defendants of each race were prosecuted. Since Y_{it} varies with time, we estimate average prosecuted outcomes in each period t: $E[Y_{it}(1)|R_i = r]$. We then measure racial gaps in prosecution that account for such differences, following Section 2.

Bounding the race-specific average outcomes if prosecuted

Figure 4 displays how average re-offence outcomes if prosecuted vary across always takers, compliers, and never takers for white and minority defendants.[36] Panels a) and c) present estimates from before the reform, and Panels b) and d) present estimates from after the reform. The first thing to note is that for both racial groups, marginally prosecuted individuals (compliers) are more likely to commit a new offence if prosecuted on average than those who are very likely to be prosecuted (always takers). One potential reason for this pattern would be if prosecutors focused their effort on individuals who were less likely to re-offend if prosecuted. The second thing to note is that the average outcomes for always takers and compliers for white defendants are uniformly lower than for minority defendants. This pattern suggests that there likely are racial differences in the average re-offence outcomes that we would see if all defendants were prosecuted.

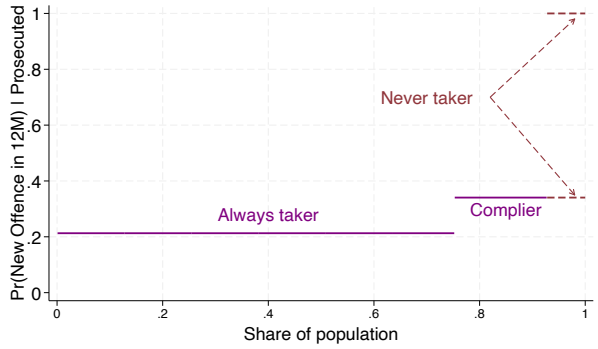
As discussed earlier, the next step assumes that the average potential outcome if prosecuted is weakly monotonic across always takers, compliers, and never takers, which implicitly places assumptions on the underlying decision-maker behavior. Here, the assumption implies that the defendants who prosecutors are least likely to prosecute (never takers) are at least as likely to re-offend if prosecuted as marginal defendants are, giving us the bounds denoted by the dashed lines in Figure 4.[37] Since never takers comprise 7-8% of the population of each racial group, we have to bound outcomes for a relatively small portion of the population. As a result, the estimated bounds on the average outcomes if everyone were prosecuted will be relatively tight. Since these

[36] As discussed in Section 2, estimating average potential outcomes by 'compliance group' is valid under IV monotonicity, which rules out the existence of 'defiers'. Here, 'defiers' would be individuals who would not be prosecuted before the reform but would be after the reform. We assess the likelihood of 'defiers' in this context by re-estimating the first-stage separately by race and baseline covariate (age bins, gender, criminal history). We find consistently large and significant reductions in prosecution similar to the estimates in Figure 2 across all race × covariate cells, suggesting that 'defiers' are unlikely to be present here (Figure A9).

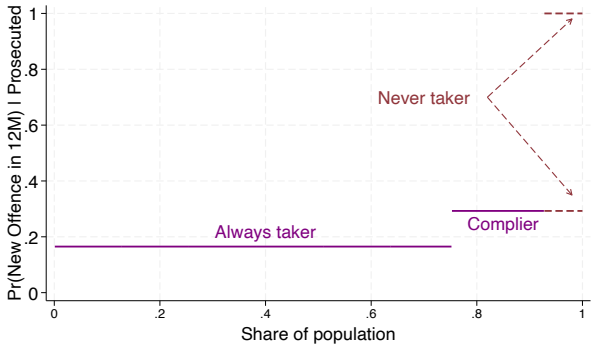
[37] This assumption could be violated if other inputs into prosecution decisions are correlated with re-offence outcomes in specific ways. E.g., if i) never takers are not prosecuted because their cases are represented by private attorneys, and ii) individuals with private representation are systematically unlikely to re-offend if prosecuted. While we do not have evidence that such violations occur in our setting, we test robustness to not assuming weak monotonicity and find similar results.

Figure 4: Average re-offence outcomes if prosecuted ($Y_i(1)$) by compliance group

(a) White defendants, pre-reform ($Z = 0$)

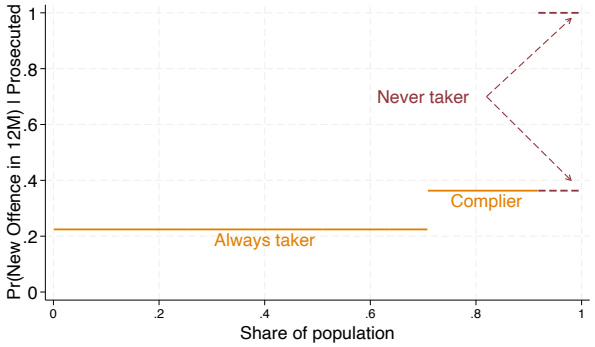


(b) White defendants, post-reform ($Z = 1$)



(c) Minority defendants, pre-reform ($Z = 0$)

(d) Minority defendants, post-reform ($Z = 1$)



Note: This figure shows the average treated outcomes for always takers, compliers, and never takers for each time period. The treatment is prosecution and the treated outcome, $Y_i(1)$, is whether an individual re-offends one year after disposition, if prosecuted. The bounds for the treated outcomes for never takers come from the assumption of weak monotonicity of average treated outcomes across compliance groups, and that $Y_i(1) \in \{0, 1\}$.

bounds are inputs into estimating the discrimination estimands (see Equation 3), we should obtain relatively tight bounds on discrimination as well.

Figure 5 estimates bounds for the average prosecuted outcome by computing a weighted average of the average outcomes for always takers, compliers, and never takers from Figure 4. We find meaningful and significant cross-race differences in the average outcomes that would be realized if all defendants were prosecuted. Before the budget reform, approximately 24–29% of white defendants would have re-offended if prosecuted, while approximately 32–37% of minority defendants would have done so. Using a bootstrapped inference procedure described in Appendix B.6, we reject the null that these bounds overlap ($p = 0.006$). After the reform, the bounds on average prosecuted outcomes are still meaningfully different (approximately 20–25% vs 27–32%), although testing the probability that they overlap is somewhat less precise ($p = 0.093$).[38] Given that re-offence outcomes if prosecuted differ meaningfully by racial group, not accounting for unobservable cross-

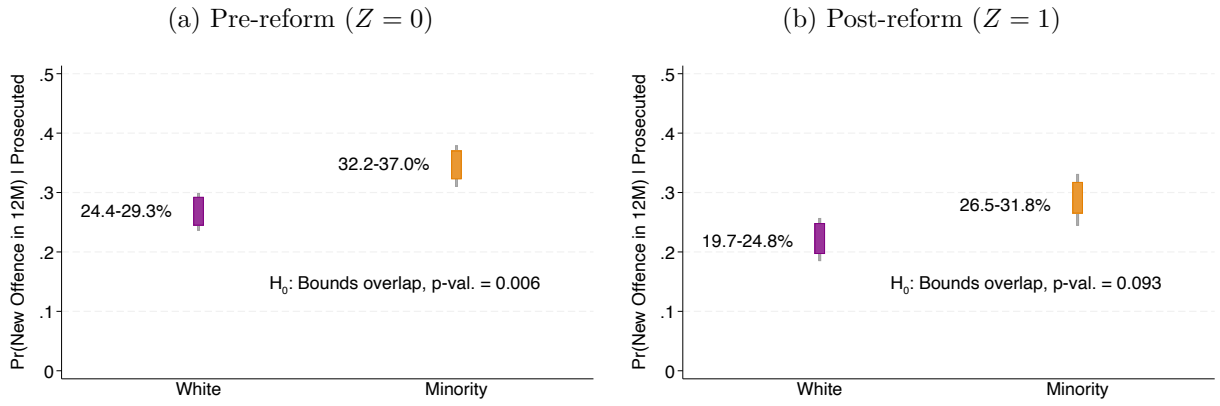
[38]These patterns are not due to our decision to define a broad minority subsample. Figure A10 disaggregates the average re-offence outcomes if prosecuted for minority defendants separately by the largest race/ethnicity subcategories. While there is variation across these subcategories, the average re-offence rates for non-Black and non-Hispanic defendants are not large outliers.

race differences would yield incorrect estimates of discrimination.

Estimating the racial gap in prosecution

Next, we estimate bounds for the racial differences in prosecution rates that condition on re-offence outcomes if prosecuted. Following Equation 3, we need, for each racial group and time period, 1) the average re-offence outcomes among prosecuted defendants, 2) the prosecution rate, and 3) the average re-offence outcomes if everyone were prosecuted. 1) and 2) are directly observed in the data, and we use bounds for 3) from Figure 5.

Figure 5: Average re-offence outcomes if prosecuted, $E[Y_{it}(1)|R_i = r]$



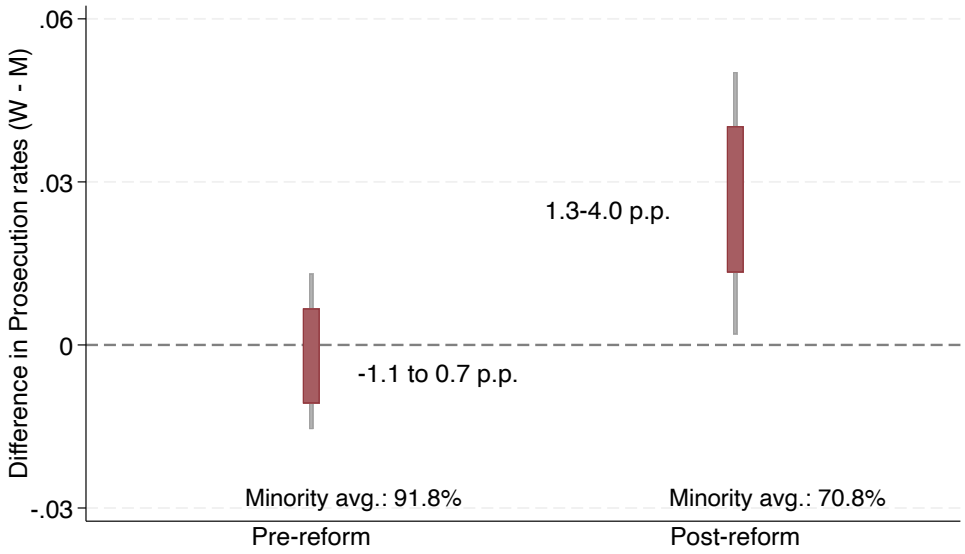
Note: This figure presents bounds on the average treated outcome obtained using the approach described in Section 2.2, separately by race and time period. The treatment is prosecution and the treated outcome, $Y_i(1)$, is whether an individual re-offends one year after disposition, if prosecuted. Confidence intervals are for the true parameter and are bootstrapped using 1,000 replications and a Bayesian bootstrap. The p-value is from a formal bootstrapped test of whether the identified sets overlap, described in Appendix B.6.

Figure 6 displays bounds on the average white–minority gap in prosecution rates conditional on re-offence outcomes if prosecuted. After accounting for racial differences in the outcomes if prosecuted, we cannot reject that white and minority defendants were prosecuted at similar rates before the budget reform. Our bounds suggest that white defendants before the reform were between 0.7 p.p. **more likely** to 1.1 p.p. **less likely** to be prosecuted compared to their minority counterparts. While these bounds suggest that prosecution in this context may not have been discriminatory before the reform, this does not speak to discrimination in other aspects of the criminal legal system or society.

This pattern changes after the reform—white defendants were 1.3–4 p.p. (1.7–5.3%) more likely than minority defendants to be prosecuted, after accounting for racial differences in the re-offence outcomes if prosecuted. Even though the budget reform reduces prosecution rates overall, the reduction is greater for minority defendants than for white defendants, even after conditioning on re-offence outcomes if prosecuted.

Alternative approaches to estimate discrimination here yield estimates outside of our estimated bounds. 'Selection-on-observable' estimates that control for age, gender, criminal history, and court fixed effects would estimate white–minority prosecution gaps of 1.3 p.p. before the budget reform

Figure 6: Racial prosecution gap conditional on prosecuted outcome



Note: This figure presents bounds on the average difference in prosecution rates in each time period, conditional on prosecuted outcomes, $Y_i(1)$, using the approach described in Section 2.2. $Y_i(1)$ is whether an individual re-offends one year after disposition, if prosecuted. Confidence intervals are for the true parameter and are bootstrapped using 1,000 replications and a Bayesian bootstrap.

and 4.5 p.p. after the reform. This alternative estimate of discrimination in prosecution before the reform is potentially incorrectly signed, and the alternative estimate of discrimination after the reform is biased upward by 13%–246%.

Robustness checks

These findings are robust to changing our empirical definitions of racial groups, prosecution, and re-offence outcomes. First we demonstrate that these patterns are not driven by our definition of the 'minority' group. Figure A11 and Figure A12 show that even after excluding Asian American & Pacific Islander defendants from the minority group, the average re-offence outcomes if prosecuted and discrimination estimates are qualitatively and quantitatively similar to our baseline. Our findings are also robust to a broader definition of prosecution that includes fine-only punishments as prosecution—Figure A13 shows racial gaps that are qualitatively similar to and overlap considerably with the baseline bounds. We also show that our baseline results remain similar when we condition on re-offending for types of offences that are likely considered high-priority by the prosecutors' office, where we use a revealed preference approach to identify high-priority offences (see Figure A14). The racial gaps in prosecution that we estimate also remain stable even if we reduce or expand the time horizon used to measure re-offending (see Panel (a) of Figure A15).

Our results are also robust to weakening key identifying assumptions. The patterns in Figure 6 are not driven by the weak monotonicity assumption imposed in Figure 4. Figure A16 relaxes that assumption and constructs the widest possible bounds by allowing never takers' outcomes to lie

between 0 and 1—i.e., allowing none or all of the never takers to re-offend if prosecuted (Manski, 1989). This exercise yields nearly identical discrimination estimates to our baseline.

Finally, Figure A17 reports estimates of discrimination when we simultaneously adjust for potential re-offence outcomes and baseline covariates (age, gender criminal history, and court fixed effects). We continue to find patterns similar to our baseline estimates, implying that covariates in this context provide little additional information not already captured by potential outcomes. These results also suggest that the unwarranted disparities we estimate are not mediated by the covariates we have access to.

3.3 Understanding drivers of the racial gap in prosecution after the reform

Our results so far document that in King County: 1) there was little evidence of discrimination in prosecution before the budget reform and that 2) even though the reform reduced overall prosecution rates, white defendants were more likely to be prosecuted than minority defendants who would experience identical re-offence outcomes if prosecuted. Next, we investigate potential factors that could be driving this result.

Since the budget reform posed significant strain on resources, the relatively higher prosecution rate for white defendants (even conditional on prosecuted outcome) could stem from cases involving minority defendants needing more resources to prosecute. Discrimination in earlier stages of the criminal legal system, such as the arrest decision or the decision to accept cases from police, could lead the cases in our court data that involve minority defendants to be backed up by weaker evidence (Goncalves and Mello, 2021; Owens and Ba, 2021; Jordan, 2024). Since cases with weak evidence would require greater resources to prosecute successfully, prosecutors may have been less likely to pursue such cases after the budget reform, when resources became scarce. Instead, prosecutors might have shifted resources to cases that they were more likely to win. As mentioned in Section 3.1, the King County Prosecutors' Office expressed concerns at the time about not being able to prosecute resource-intensive and time-consuming cases.

We investigate this explanation by using our approach to estimate discrimination in two subsamples that vary in terms of average case quality. We use a data-driven procedure to classify offence types into two bins based on whether the share of charges successfully punished was above or below the median successful punishment rate across offence types, using pre-reform King County data from September 2004 to September 2010. The logic is that we should see a high conversion rate of charges into punishments for offences where the average case is typically backed up by high quality evidence. We refer to such offences as "high quality" cases, which include drug, driving under the influence, property, prostitution, and weapons violations. On the other hand, arrests for traffic, violent, and 'other' offences have a relatively low share of charges that are punished, and we refer to these types of cases as "low quality".[39] As Table 2 shows, before the budget reform, cases

[39] One potential reason why quality differs across these offence categories is the objectivity of evidence. For example, driving under the influence cases might be backed up by blood alcohol content tests, which are relatively objective and verifiable, while cases involving violent offences may be more likely to rely on subjective witness statements.

involving minority defendants were more likely to be lower "quality" than cases involving white defendants.

While other factors may also vary between these two categories (e.g., amount of resources needed to prosecute), if the budget cut forced prosecutors to re-allocate resources based on case quality, and minority defendants' cases tended to be lower quality, we should expect to see a more positive post-budget cut gap within the "low quality" subset.

Table 2: Distribution of case "quality" by defendant racial group

	Share of cases that are:	
	High quality	Low quality
White	0.24	0.76
Minority	0.20	0.80

Note: Sample includes all criminal cases disposed in the King County District Courts between October 2008 and September 2010 (i.e., in the analysis sample but before the budget reform). 'High quality'/'low quality' offences are those with an above/below median share of charges that result in any punishment in this time period.

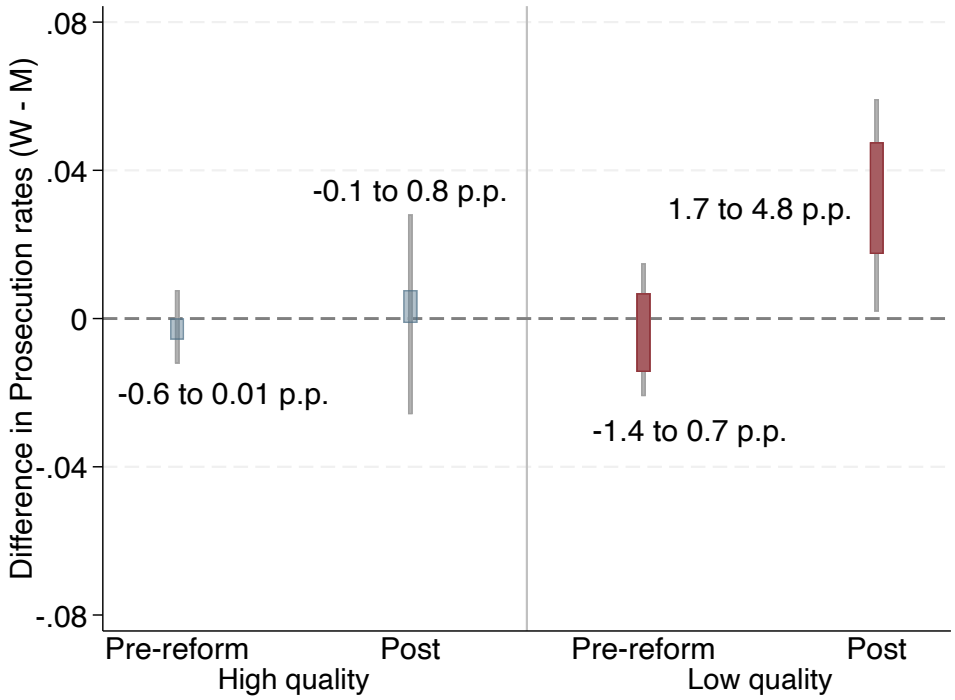
Figure 7 displays our discrimination estimates separately for "high quality" and "low quality" cases. We see that our findings of discrimination after the reform are driven by the "low quality" cases. In this subsample, white defendants after the reform were 1.7–4.8 p.p.more likely than minority defendants to be prosecuted, after accounting for racial differences in the re-offence outcomes if prosecuted. In contrast, we see a post-period gap between –0.1 to 0.8 p.p. among the "low quality" subset of cases, which is much smaller and lies outside of the post-period "high quality" bounds.[40] However, we cannot reject that these bounds overlap due to the reduction in power from splitting our data into two subsamples.

While this exercise only provides suggestive evidence, the divergence in discrimination patterns between the "high quality" and "low quality" subsamples and the patterns in Table 2 suggest that minority defendants' cases might be lower quality, and that such cases are dropped due to the budget reform.[41] Taken seriously, this finding would imply that prosecutors were pursuing most cases before the reform, and passing through any disparities generated in stages in the criminal legal system before the prosecution decision, e.g., police arrest or prosecutor charging decisions (Harrington and Shaffer, 2024). However, in shifting their focus to high quality cases, prosecutors may have offset some of those disparities from prior stages. While this behavior is consistent with

[40] We also use the approach in Appendix B.1 to simultaneously condition on whether defendants would re-offend if prosecuted as well as whether the case would be successful if prosecuted, where we consider a case successful if any charge received a sentence. Panels a) through d) of Figure A18 display racial gaps for each combination of these binary potential outcomes. While the bounds are generally quite wide, we see more positive racial gaps among the subset of cases that would not have been successful if prosecuted (top row), which is consistent with the patterns in Figure 7.

[41] Some cases in the data are dismissed without charges ever being recorded and we consider these cases as "low quality" in our baseline classification. If we exclude such cases from the exercise, we see similar shares of "low quality" cases for white defendants (76%) and minority defendants (79%) as in Table 2 and we obtain similar but less precise conclusions as in Figure 7 (see Figure A19).

Figure 7: Racial prosecution gap, by proxy for case quality



Note: This figure presents bounds on the average difference in prosecution rates in each time period, conditional on prosecuted outcomes, $Y_i(1)$, using the approach described in Section 2.2. $Y_i(1)$ is whether an individual re-offends one year after disposition, if prosecuted. 'High quality'/'low quality' offences are those with an above/below median share of charges that result in any punishment using pre-reform data. Confidence intervals are for the true parameter and are bootstrapped using 1,000 replications and a Bayesian bootstrap.

prior work showing how prosecutors can use discretion to attenuate prior discrimination (Harrington and Shaffer, 2023; Jordan, 2024), it contrasts with how other agents in the criminal legal system react when fiscally-constrained—e.g., when budget deficits bind, police alter their search behavior in ways that increase rather than reduce racial disparities (Makowsky, Stratmann, and Tabarrok, 2019).

3.4 Conditioning on the outcome if dismissed or the treatment effect

We now consider alternate definitions of racial discrimination in prosecution that condition on the outcome if **dismissed** or the **treatment effect** of prosecution. Starting with the outcome if dismissed, this definition can be interpreted as an estimate of discrimination that holds fixed a notion of the baseline "risk" that an arrested individual might re-offend. As discussed in Appendix B.1 (Equation 5), estimating this quantity first requires bounding outcomes if dismissed for always takers (who are always prosecuted). Given that always takers are approximately 80% of the population in this setting (see Figure 4), we should expect the bounds on average dismissed outcomes and discrimination to be wider here.

Again, given the time-varying nature of the potential outcomes, we must purge the time trends

in potential re-offence outcomes if dismissed ($Y_{it}(0)$) in King County using trends in $Y_{it}(0)$ in adjacent counties (see Appendix B.5 for details). We make a similar parallel trends assumption as we previously did, but now for the re-offence outcomes if **dismissed**. As described in **A3**, the assumption requires parallel trends to hold between compliers and never takers (not always takers). We use this assumption to account for time trends in the average dismissed outcomes for compliers and never takers, and construct bounds for always takers' dismissed outcomes.[42]

A3 Without the reform, re-offence outcomes if dismissed, $Y_{it}(0)$, would trend similarly for never takers & compliers and is independent of county.

Using the time trend adjustment, Figure A20 presents estimates of the average re-offence rates if all individuals were dismissed, separately by race and time period. We find suggestive evidence that minority defendants are more likely than white defendants to commit a new offence if dismissed, but we cannot reject that the bounds for the two racial groups overlap. For example, our estimates suggest that before the reform, 4.3%–15.4% of white defendants would commit a new offence if all were dismissed, while 6.5%–20.1% of minority defendants would commit a new offence if dismissed.

We use the bounds on the race-specific average outcomes if dismissed to compute racial gaps in prosecution, conditional on the outcome if dismissed. Figure A21 shows that similar to our baseline results, we cannot reject that there is no discrimination in prosecution prior to the reform, and that white defendants are 1.0–8.2 p.p. (1.4–11.6%) more likely to be prosecuted after the reform. Given the high prosecution rates in this setting, this is a scenario where the weak monotonicity assumption dramatically tightens the bounds relative to alternative approaches that use weaker assumptions. For example, 'worst case' bounds for discrimination conditional on the outcome if dismissed after the reform range between –17.9 to 38.8 p.p. (see Figure A22).

As with our baseline results, the findings here are also robust to altering the time horizon used to measure re-offending (see Panel (b) of Figure A15) and simultaneously controlling for baseline covariates (see Panel (b) of Figure A17). We also redo the exercise estimating racial gaps in prosecution separately for potentially "high quality" and "low quality" cases, this time conditioning on re-offence outcomes if dismissed. The patterns are qualitatively similar to what we see when conditioning on re-offence outcomes if prosecuted, but the bounds are wider and less precise (see Figure A23).

We also implement the tools described in Appendix B.1 to measure racial gaps in prosecution that are conditional on both the outcome if prosecuted and dismissed, to understand if the racial gaps are concentrated among certain values of the treatment effect of prosecution. Panels a) through d) of Figure A24 display racial gaps for each combination of the binary potential outcomes. Panel b) depicts racial gaps among those for whom the treatment effect of prosecution is positive

[42] Similar to the previous validation exercises, we test for differential pre-trends in the outcomes of dismissed individuals across county and covariate subgroups. Figure A29 shows no evidence of differential pre-trends in re-offence outcomes if dismissed across counties but within various demographic subgroups. Figure A30 finds limited evidence of differential pre-trends in re-offence outcomes if dismissed within counties but across subgroups. Almost all estimates are statistically indistinguishable from zero, which we interpret as evidence that **A3** is not grossly violated.

$(Y(1) = 1, Y(0) = 0)$, while Panel c) displays racial gaps for those with a negative treatment effect $(Y(1) = 0, Y(0) = 1)$. The racial gaps in Panels a) and c) are for defendants for whom prosecution has no impact on re-offending within one year. The bounds are generally too wide to draw any serious conclusions, which is driven by the width of the bounds for the average outcome if dismissed discussed above.

3.5 Summary

Our analysis presents the first evidence of racial discrimination in misdemeanor prosecution that directly accounts for meaningful unobservable differences across groups. We find that after a budget reform that limited prosecutor capacity and reduced overall prosecution rates, white defendants were more likely to be prosecuted than minority defendants with similar potential re-offence outcomes. Digging deeper, our evidence suggests that prosecutors seemed to be dismissing low quality and potentially resource-intensive cases, which were more prevalent among minority defendants. As a result, this behavior may have attenuated discrimination from stages of the criminal legal system before the prosecution decision that contributed to the case quality gaps in the first place, e.g., police arrest or prosecutor charging decisions.

4 Conclusion

This paper shows how to use a natural experiment that generates a binary instrumental variable (IV) to estimate bounds or point estimates of discrimination conditional on potential outcomes, with extensions to condition on treatment effects or multiple distinct potential outcomes. We do this by combining the shifts in treatment rates and outcomes induced by the binary IV with behavioral assumptions on the relationship between selection into treatment and average potential outcomes from the marginal treatment effects literature. We also discuss how to apply our approach when a natural experiment generates a conditionally-random binary IV, as in the case of regression discontinuity or difference-in-difference (DiD) designs. In doing so, we add to the DiD-IV literature by showing how to use selection into being an always taker, complier, or never taker to estimate average potential outcomes with DiD variation.

Our approach does not require random assignment to decision-makers, or even that the key decision-makers are observed in the data, in contrast to existing discrimination estimation methods. Thus, our approach allows researchers to study discrimination in high-stakes settings where individuals are typically not randomly assigned to the decision-makers—for example, managers' decisions to promote workers or teachers' decisions to suspend students.

We implement our approach to provide novel evidence on racial discrimination in misdemeanor prosecution, the most common form of contact with criminal courts. We use a DiD-IV strategy generated by an unexpected budget cut that affected the King County Prosecutors' Office but did not affect adjacent counties. Adjusting for racial differences in potential re-offence outcomes, we find no evidence of discrimination in prosecution before the budget cut. After the cut, minority

defendants were less likely to be prosecuted than white defendants with identical potential re-offence outcomes, even though overall prosecution rates fell. We find suggestive evidence of prosecutors responding to fiscal constraints by focusing on easy cases and offsetting disparities from prior stages of the criminal legal system.

While we discuss measuring discrimination conditional on treatment effects or multiple potential outcomes, the resulting bounds tend to be wide in practice. Future methodological steps might involve narrowing the resulting bounds, perhaps by imposing structural restrictions or incorporating insights from tools that estimate average population outcomes while accounting for two dimensions of unobservable heterogeneity (Dutz et al., 2021). Fruitful next steps for the empirical analysis of discrimination in prosecution might further assess the mechanisms underlying how discrimination is affected by changes in resource constraints. In particular, combining our approach with data tracking individual events within each court case would allow us to better understand which prosecutorial actions amplify versus offset any pre-existing disparities.

References

Abrams, David S., Marianne Bertrand, and Sendhil Mullainathan. 2012. "Do Judges Vary in Their Treatment of Race?" *The Journal of Legal Studies* 41 (2): 347–383.

Agan, Amanda, Jennifer L Doleac, and Anna Harvey. 2023. "Misdemeanor Prosecution." *The Quarterly Journal of Economics* 138 (3): 1453–1505.

Agan, Amanda Y. 2024. "Racial Disparities in the Criminal Legal System: Shadows of Doubt and Beyond." *Journal of Economic Literature* .

Aigner, Dennis J., and Glen G. Cain. 1977. "Statistical Theories of Discrimination in Labor Markets." *Industrial and Labor Relations Review* 30 (2): 175–187.

Amaral, Francesca A., Aurélie Ouss, and Dalila I. Ozier. 2025. "Prosecutor-driven reform and racial disparities." *Criminology & Public Policy* .

Angelova, Victoria, Will S. Dobbie, and Crystal Yang. 2023. "Algorithmic Recommendations and Human Discretion."

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91 (434): 444–455.

Angrist, Joshua D., and Miikka Rokkanen. 2015. "Wanna Get Away? Regression Discontinuity Estimation of Exam School Effects Away From the Cutoff." *Journal of the American Statistical Association* 110 (512): 1331–1344.

Anwar, Shamena, Patrick Bayer, and Randi Hjalmarsson. 2012. "The Impact of Jury Race in Criminal Trials*." *The Quarterly Journal of Economics* 127 (2): 1017–1055.

Anwar, Shamena, and Hanming Fang. 2006. "An Alternative Test of Racial Prejudice in Motor Vehicle Searches: Theory and Evidence." *THE AMERICAN ECONOMIC REVIEW* 96 (1).

Arnold, David, Will Dobbie, and Peter Hull. 2022. "Measuring Racial Discrimination in Bail Decisions." *American Economic Review* 112 (9): 2992–3038.

Arnold, David, Will Dobbie, and Crystal S. Yang. 2018. "Racial Bias in Bail Decisions." *The Quarterly Journal of Economics* 133 (4): 1885–1932.

Arrow, Kenneth J. 1973. "THE THEORY OF DISCRIMINATION." In *THE THEORY OF DISCRIMINATION*, 1–33. Princeton University Press.

Ayres, Ian. 2010. "Testing for Discrimination and the Problem of "Included Variable Bias"." .

Ballotpedia. 2010. "King County Public Safety Sales Tax Increase (November 2010)."

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

Becker, Gary Stanley. 1957. *The Economics of Discrimination*. University of Chicago Press.

Berne, Jordy, Brian Jacob, Christina Weiland, and Katharine Strunk. 2023. "Staying Back to Catch Up? Impacts of Michigan's Third Grade Retention Law on Children's Educational Trajectories and Academic Skills." Tech. rep.

Bertrand, Marianne, and Sendhil Mullainathan. 2004. "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." *American Economic Review* 94 (4): 991–1013.

Bharadwaj, Prashant, Rahul Deb, and Ludovic Renou. 2024. "Statistical Discrimination and the Distribution of Wages."

Bohren, J. Aislinn, Kareem Haggag, Alex Imas, and Devin G. Pope. 2019. "Inaccurate Statistical Discrimination: An Identification Problem."

Bohren, J. Aislinn, Peter Hull, and Alex Imas. 2022. "Systemic Discrimination: Theory and Measurement." Working Paper 29820, National Bureau of Economic Research.

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer. 2016. "Stereotypes*." *The Quarterly Journal of Economics* 131 (4): 1753–1794.

Brinch, Christian N., Magne Mogstad, and Matthew Wiswall. 2017. "Beyond LATE with a Discrete Instrument." *Journal of Political Economy* 125 (4): 985–1039.

Buch, Jason, and Joy Borkholder. 2020. "Report: Washington State Patrol singles out Native American drivers."

Bushway, Shawn, Andrew Jordan, Derek Neal, and Steven Raphael. 2025. "Understanding Racial Disparities in Criminal Court Outcomes."

Campbell, Romaine A. 2023. "What Does Federal Oversight Do to Policing and Public Safety? Evidence from Seattle." .

Canay, Ivan A, Magne Mogstad, and Jack Mountjoy. 2024. "On the Use of Outcome Tests for Detecting Bias in Decision Making." *The Review of Economic Studies* 91 (4): 2135–2167.

Cattaneo, Matias D, Nicolas Idrobo, and Rocıo Titiunik. 2024. "A Practical Introduction to Regression Discontinuity Designs: Volume II." *Cambridge Elements: Quantitative and Computational Methods for Social Science* .

Cattaneo, Matias D., Michael Jansson, and Xinwei Ma. 2018. "Manipulation Testing Based on Density Discontinuity." *The Stata Journal* 18 (1): 234–261.

Cattaneo, Matias D., Luke Keele, Rocío Titiunik, and Gonzalo Vazquez-Bare. 2021. "Extrapolating Treatment Effects in Multi-Cutoff Regression Discontinuity Designs." *Journal of the American Statistical Association* 116 (536): 1941–1952.

Cengiz, Doruk, Arindrajit Dube, Attila Lindner, and Ben Zipperer. 2019. "The Effect of Minimum Wages on Low-Wage Jobs*." *The Quarterly Journal of Economics* 134 (3): 1405–1454.

Charles, Kerwin Kofi, and Jonathan Guryan. 2011. "Studying Discrimination: Fundamental Challenges and Recent Progress." *Annual Review of Economics* 3 (Volume 3, 2011): 479–511.

Constantine, Dow. 2010. "Executive proposed budget to include nearly $4 million in cuts to services provided by Prosecutor - King County, Washington."

Davis, Angela J. 2014. "IN SEARCH OF RACIAL JUSTICE: THE ROLE OF THE PROSECUTOR." *LEGISLATION AND PUBLIC POLICY* 16.

De Chaisemartin, C., and X. D'Haultfœuille. 2018. "Fuzzy Differences-in-Differences." *The Review of Economic Studies* 85 (2 (303)): 999–1028.

Donahue, Allison R. 2023. "As lawmakers revamp 3rd grade reading law, advocates say dyslexia supports are needed • Michigan Advance."

Dutz, Deniz, Ingrid Huitfeldt, Santiago Lacouture, Magne Mogstad, Alexander Torgovitsky, and Winnie Van Dijk. 2021. "Selection in Surveys: Using Randomized Incentives to Detect and Account for Nonresponse Bias." Tech. Rep. w29549, National Bureau of Economic Research, Cambridge, MA.

Ervin, Keith. 2010. "Prosecutor Dan Satterberg warns of fallout from potential layoffs." *Seattle Times* .

French, Ron. 2019. "Michigan is investing heavily in early reading. So far, it's not working. | Bridge Michigan."

Gershowitz, Adam M, and Laura R Killinger. 2011. "The State (Never) Rests: How Excessive Prosecutorial Caseloads Harm Criminal Defendants." *NORTHWESTERN UNIVERSITY LAW REVIEW* .

Goldin, Claudia, and Cecilia Rouse. 2000. "Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians." *The American Economic Review* 90 (4): 715–741.

Goncalves, Felipe, and Steven Mello. 2021. "A Few Bad Apples? Racial Bias in Policing." *American Economic Review* 111 (5): 1406–1441.

Grossman, Joshua, Julian Nyarko, and Sharad Goel. 2024. "Reconciling Legal and Empirical Conceptions of Disparate Impact: An Analysis of Police Stops Across California." *Journal of Law and Empirical Analysis* 1 (1): 2755323X241243168.

Harrington, Emma, and Hannah Shaffer. 2023. "Brokers of Bias."

———. 2024. "Statistical Discrimination in Sequential Systems: Prosecutors' Response to Police." Working Paper.

Heckman, James J, and Edward J Vytlacil. 2000. "Local Instrumental Variables."

Hellerstein, Judith K., David Neumark, and Kenneth R. Troske. 1999. "Wages, Productivity, and Worker Characteristics: Evidence from Plant-Level Production Functions and Wage Equations." *Journal of Labor Economics* 17 (3): 409–446.

———. 2002. "Market Forces and Sex Discrimination." *The Journal of Human Resources* 37 (2): 353–380.

Hu, Cathy, and Sino Esthappan. 2017. "Asian Americans and Pacific Islanders, a missing minority in criminal justice data | Urban Institute."

Hull, Peter. 2021. "What Marginal Outcome Tests Can Tell Us About Racially Biased Decision-Making."

Hwang, NaYoung, and Cory Koedel. 2022. "Holding Back to Move Forward: The Effects of Retention in the Third Grade on Student Outcomes." Tech. rep., Annenberg Institute at Brown University.

Hössjer, Ola, and Arvid Sjölander. 2022. "Sharp lower and upper bounds for the covariance of bounded random variables." *Statistics & Probability Letters* 182: 109323.

Imbens, Guido W., and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62 (2): 467–475.

Imbens, Guido W., and Charles F. Manski. 2004. "Confidence Intervals for Partially Identified Parameters." *Econometrica* 72 (6): 1845–1857.

Imbens, Guido W., and Donald B. Rubin. 1997. "Estimating Outcome Distributions for Compliers in Instrumental Variables Models." *The Review of Economic Studies* 64 (4): 555–574.

Jacob, Brian A., and Lars Lefgren. 2004. "Remedial Education and Student Achievement: A Regression-Discontinuity Analysis." *The Review of Economics and Statistics* 86 (1): 226–244.

Jordan, Andrew. 2024. "Racial Patterns in Approval of Felony Charges."

———. 2025. "Relaxing Charging Restrictions in Narcotics Cases."

Kaplan, Jacob. 2023. "Jacob Kaplan's Concatenated Files: Uniform Crime Reporting Program Data: Law Enforcement Officers Killed and Assaulted (LEOKA) 1960-2021."

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. "Human Decisions and Machine Predictions*." *The Quarterly Journal of Economics* 133 (1): 237–293.

Knowles, John, Nicola Persico, and Petra Todd. 2001. "Racial Bias in Motor Vehicle Searches: Theory and Evidence." *Journal of Political Economy* 109 (1): 203–229.

Kowalski, Amanda E. 2023a. "Behaviour within a Clinical Trial and Implications for Mammography Guidelines." *The Review of Economic Studies* 90 (1): 432–462.

Kowalski, Amanda E. 2023b. "Reconciling Seemingly Contradictory Results from the Oregon Health Insurance Experiment and the Massachusetts Health Reform." *Review of Economics and Statistics* 105 (3): 646–664.

Kutateladze, Besiki Luka, and Nancy R Andiloro. 2014. "Prosecution and Racial Justice in New York County – Technical Report." .

Leasure, Peter. 2019. "Misdemeanor Records and Employment Outcomes: An Experimental Study." *Crime & Delinquency* 65 (13): 1850–1872.

Makowsky, Michael D., Thomas Stratmann, and Alex Tabarrok. 2019. "To Serve and Collect: The Fiscal and Racial Determinants of Law Enforcement." *The Journal of Legal Studies* 48 (1): 189–216.

Malott, Samantha. 2024. "Hidden health disparities of Native Hawaiian, Pacific Islander communities."

Manski, Charles F. 1989. "Anatomy of the Selection Problem." *The Journal of Human Resources* 24 (3): 343–360.

Marx, Philip. 2022. "An Absolute Test of Racial Prejudice." *The Journal of Law, Economics, and Organization* 38 (1): 42–91.

Michigan Department of Education. 2023. "Interpretive Guide to M-STEP Reports."

Mogstad, Magne, Andres Santos, and Alexander Torgovitsky. 2018. "Using Instrumental Variables for Inference About Policy Relevant Treatment Parameters." *Econometrica* 86 (5): 1589–1619.

Mueller-Smith, Michael, and Kevin T. Schnepel. 2021. "Diversion in the Criminal Justice System." *The Review of Economic Studies* 88 (2): 883–936.

Owens, Emily, and Bocar Ba. 2021. "The Economics of Policing and Public Safety." *Journal of Economic Perspectives* 35 (4): 3–28.

Phelps, Edmund S. 1972. "The Statistical Theory of Racism and Sexism." *The American Economic Review* 62 (4): 659–661.

Povich, Elaine S. 2023. "Debate over holding back third graders roils state legislatures • Michigan Advance."

Rehavi, M. Marit, and Sonja B. Starr. 2014. "Racial Disparity in Federal Criminal Sentences." *Journal of Political Economy* 122 (6): 1320–1354.

Ricks, Michael David. 2022. "Strategic Selection Around Kindergarten Recommendations." .

Rose, Evan K. 2021. "Who Gets a Second Chance? Effectiveness and Equity in Supervision of Criminal Offenders*." *The Quarterly Journal of Economics* 136 (2): 1199–1253.

Ross, Stephen L. 2005. "The Continuing Practice and Impact of Discrimination." *Economics Working Papers* .

Ross, Stephen L., Margery Austin Turner, Erin Godfrey, and Robin R. Smith. 2008. "Mortgage lending in Chicago and Los Angeles: A paired testing study of the pre-application process." *Journal of Urban Economics* 63 (3): 902–919.

Rubin, Donald B. 1974. "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of Educational Psychology* 66 (5): 688–701.

———. 1981. "The Bayesian Bootstrap." *The Annals of Statistics* 9 (1): 130–134.

Sloan, CarlyWill. 2022. "Do Prosecutor and Defendant Race Pairings Matter? Evidence from Random Assignment."

Spohn, Cassia, John Gruhl, and Susan Welch. 1987. "The Impact of the Ethnicity and Gender of Defendants on the Decision to Reject or Dismiss Felony Charges*." *Criminology* 25 (1): 175–192.

Stevenson, Megan T., and Sandra G. Mayson. 2018. "The Scale of Misdemeanor Justice."

Tuttle, Cody. 2023. "Racial Disparities in Federal Sentencing: Evidence from Drug Mandatory Minimums." *SSRN Electronic Journal* .

Vytlacil, Edward. 2002. "Independence, Monotonicity, and Latent Index Models: An Equivalence Result." *Econometrica* 70 (1): 331–341.

Westall, John, Tara Kilbride, Andrew Utter, and Katharine O Strunk. 2022a. "2022 Preliminary Read by Grade Three Retention Estimates." .

Westall, John, Andrew Utter, Tara Kilbride, and Katharine O Strunk. 2022b. "Read by Grade Three Law Initial Retention Decisions." .

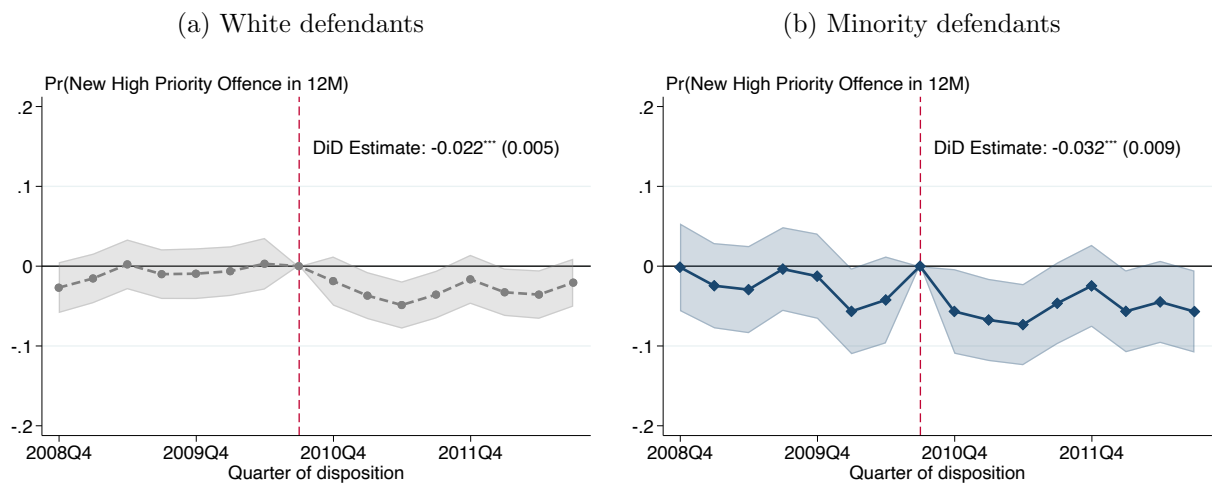
Westall, John, Andrew Utter, and Katharine O. Strunk. 2023. "Following the Letter of the Law: 2020-21 Retention Outcomes Under Michigan's Read by Grade Three Law." .

Yuan, Andy, and Spencer Cooper. 2022. "Prosecutorial Incentives and Outcome Disparities." Tech. rep.

Appendix A Additional results: Misdemeanor prosecution

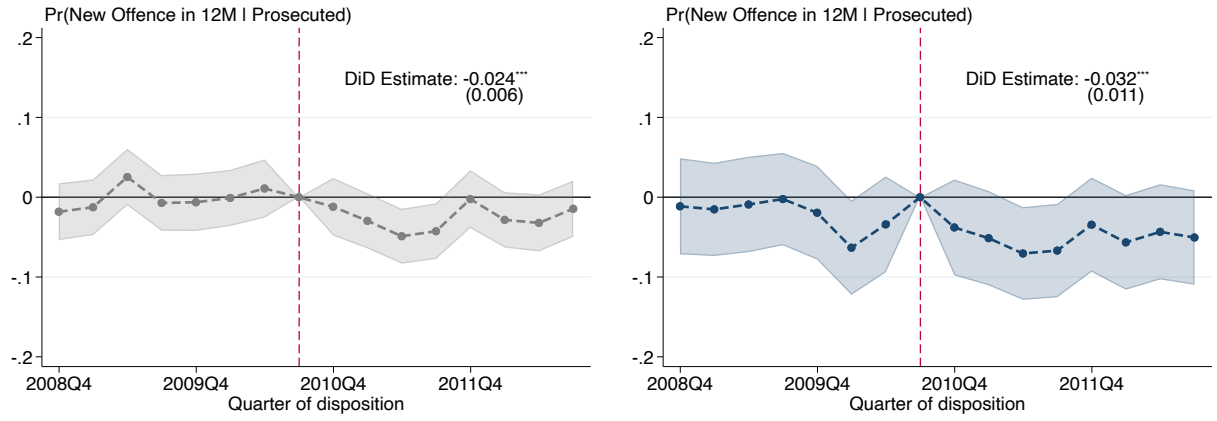
A.1 Robustness checks

Figure A1: Impact of King County budget reform on 'high-priority' re-offence within one year



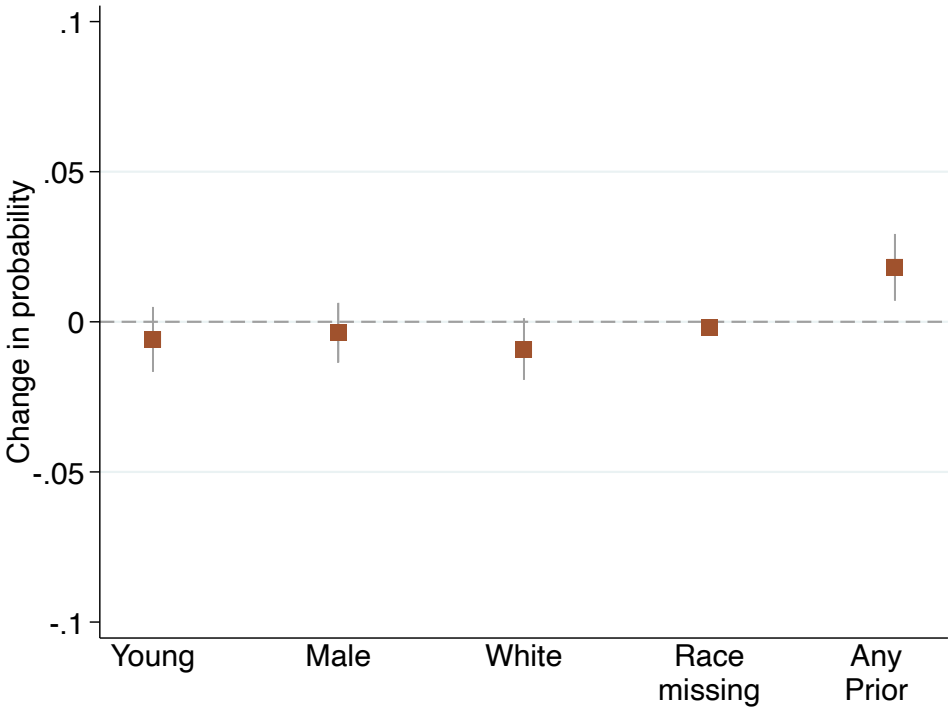
Note: Each Panel presents event study estimates investigating the impact of the King County budget reform. The re-offence outcome includes new 'high-priority' offences committed anywhere in Washington State. 'High-priority' offences are those that are not associated with charges that were commonly dismissed in the 2 quarters after the budget reform. Sample includes all misdemeanor defendants, as described in Table 1. 'DiD Estimate' pools the coefficients on relative time indicators and estimates $Y_{igt} = \alpha + \delta_1 \mathbb{I}[\text{King County}] + \delta_2 Post_i + \beta^{DD} \mathbb{I}[\text{King County}] \times Post_i + \epsilon_{igt}$, where $Post_i = 1$ if the case is disposed on or after September 28, 2010, when the budget reform was announced. 95% confidence intervals are constructed using heteroscedasticity-robust standard errors.

Figure A2: Impact of King County budget reform on re-offence within one year, only prosecuted defendants



Note: Each Panel is identical to Figure 3, except the sample only includes prosecuted defendants.

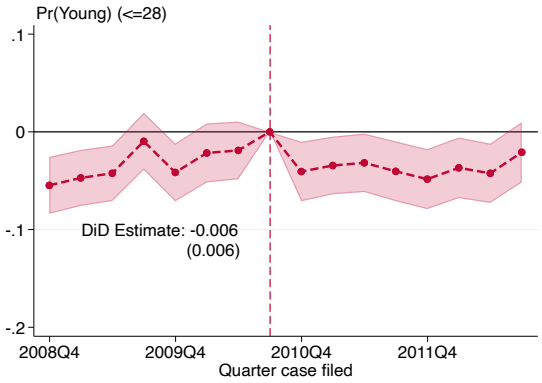
Figure A3: Testing for changes in observable characteristics of cases filed



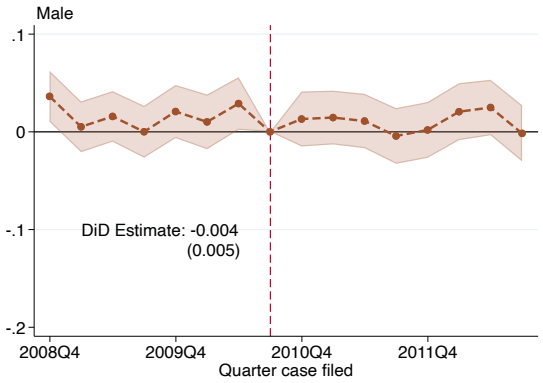
Note: Each square is β^{DD} from $X_{igt} = \alpha + \delta_1 \mathbb{I}[\text{King County}] + \delta_2 Post_i + \beta^{DD} \mathbb{I}[\text{King County}] \times Post_i + \epsilon_{igt}$, where $Post_i = 1$ if the case is filed on or after September 28, 2010, when the budget reform was announced. X_{igt} is the relevant baseline characteristic. 'Young' defendants are those who $\leq$ 28 years old at disposition and 'Any Prior' is an indicator for whether an individual has ever been previously charged with an offence in Washington State. 95% confidence intervals are constructed using heteroscedasticity-robust standard errors.

Figure A4: Impact of budget reform on caseload characteristics

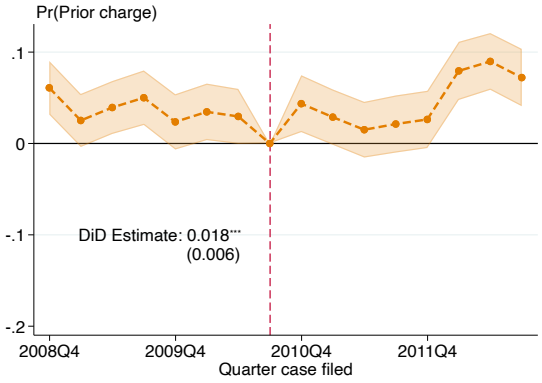
(a) Age $\leq$ 28



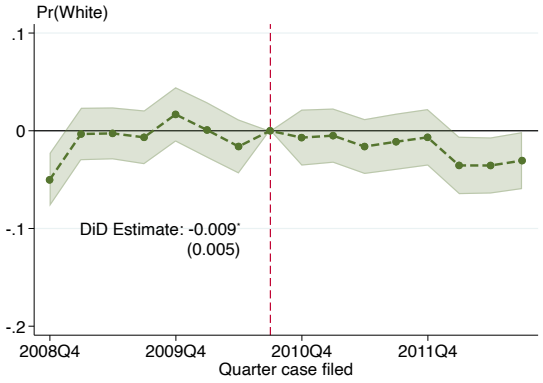
(b) Male



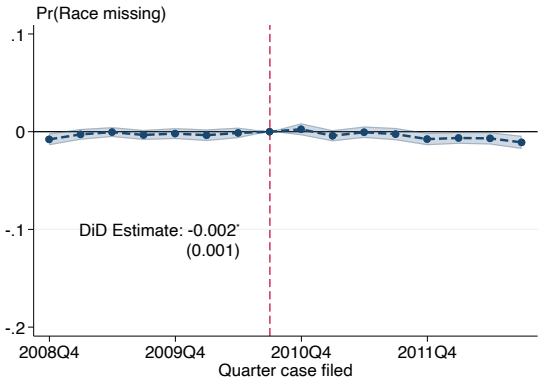
(c) Any prior charge



(d) White (Non-Hispanic)



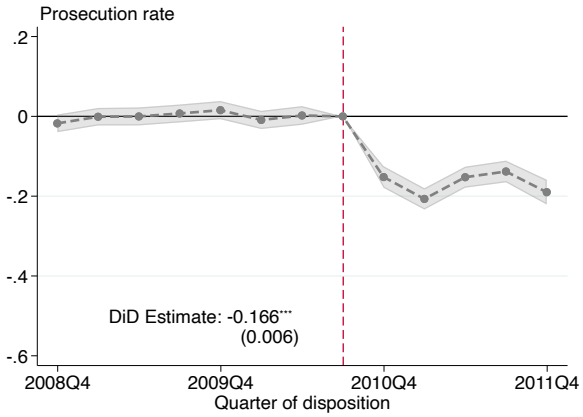
(e) Missing race/ethnicity



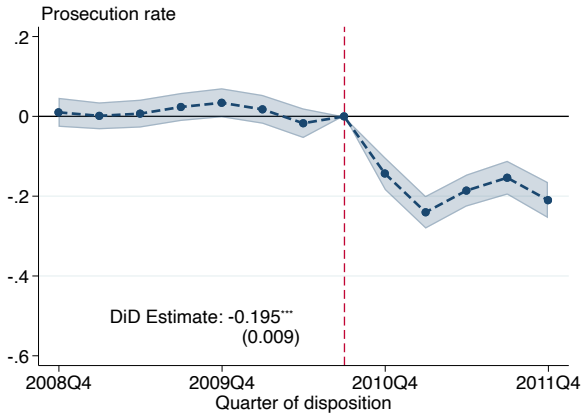
Note: Each Panel presents event study estimates investigating the impact of the King County budget reform. 'DiD estimate' is β^{DD} from $X_{igt} = \alpha + \delta_1 \mathbb{I}[\text{King County}] + \delta_2 Post_i + \beta^{DD} \mathbb{I}[\text{King County}] \times Post_i + \epsilon_{igt}$, where $Post_i = 1$ if the case is filed on or after September 28, 2010, when the budget reform was announced. X_{igt} denotes a baseline characteristic. 'Young' defendants are those who $\leq$ 28 years old at disposition and 'Any Prior' is an indicator for whether an individual has been previously charged with an offence in Washington. 95% confidence intervals constructed with heteroscedasticity-robust standard errors.

Figure A5: Robustness of first stage and reduced form to excluding post-SPD investigation period

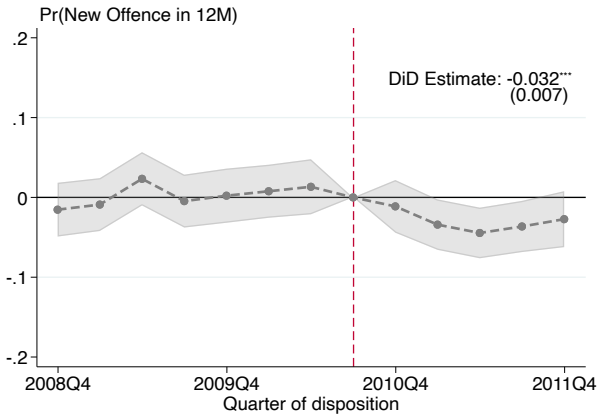
(a) First stage, White defendants



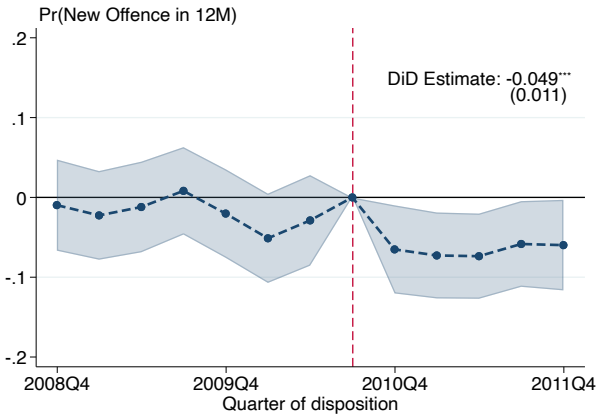
(b) First stage, Minority defendants



(c) Reduced form, White defendants



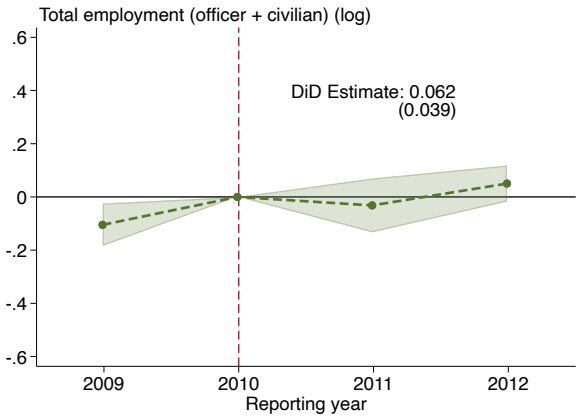
(d) Reduced form, Minority defendants



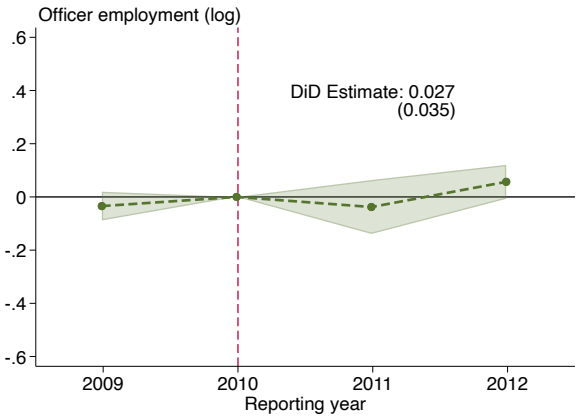
Note: This reproduces Figure 2 (Panels (a) and (b)) and Figure 3 (Panels (c) and (d)), excluding cases disposed after December 11, 2011.

Figure A6: Impact of budget reform on police employment

(a) Total employment



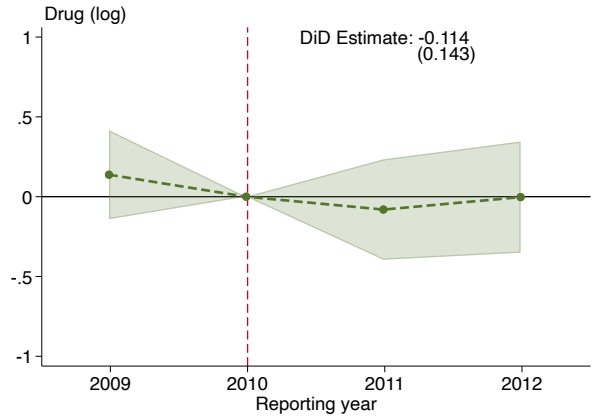
(b) Officer employment



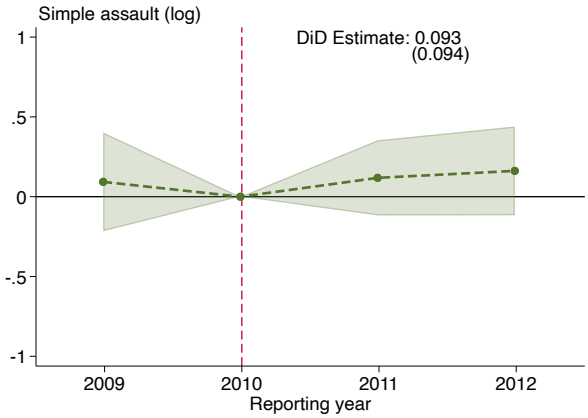
Note: Estimates generated using annual average employment from the Law Enforcement Officers Killed and Assaulted (LEOKA) data (Kaplan, 2023). This specification is similar to those used to estimate the first stage and reduced form except for the inclusion of originating agency (ORI) fixed effects and usage of the average pre-reform county-level population as weights. ORIs in areas with an annual population average less than 1,000 or with zero employment counts throughout the sample are excluded. 95% confidence intervals are constructed using standard errors clustered at the ORI level.

Figure A7: Impact of budget reform on arrests

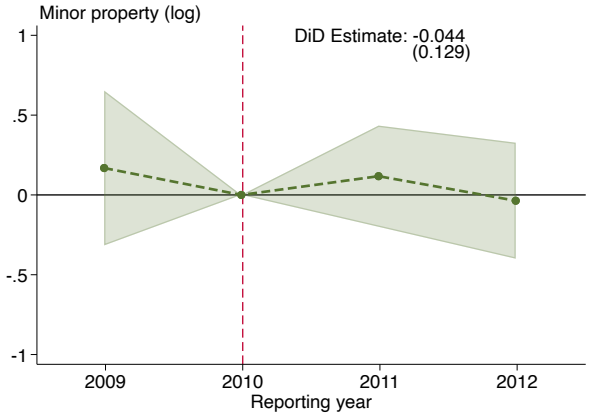
(a) Drug (possession +sale)



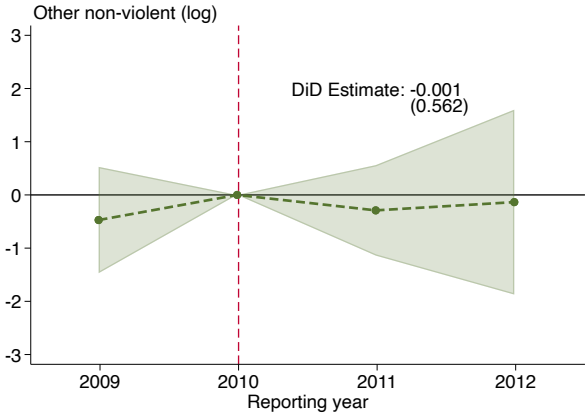
(b) Simple assault



(c) Minor property



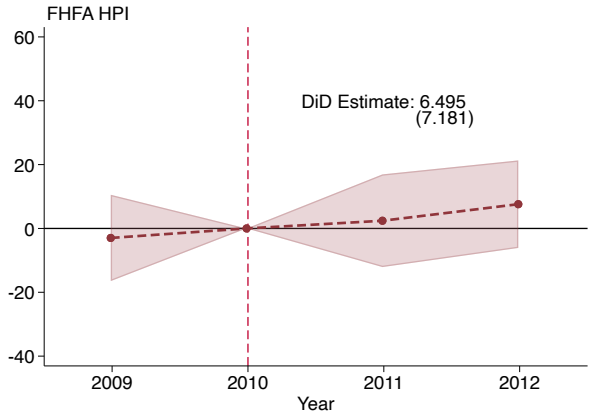
(d) Minor other arrests



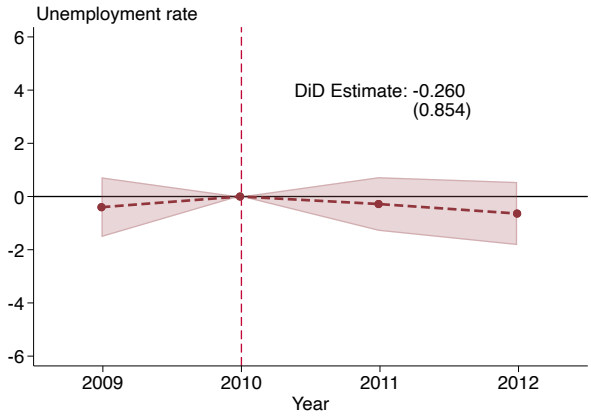
Note: Estimates generated using annual number of arrests from Uniform Crime Report (UCR) data (Kaplan, 2023). 'Minor property' arrests include arrests for stolen property, fraud, forgery, and theft. This specification is similar to those used to estimate the first stage and reduced form except for the inclusion of originating agency (ORI) fixed effects and usage of the average pre-reform county-level population as weights. ORIs in areas with an annual population average less than 1,000, with limited time coverage between 2008-13 or with zero arrest counts in a given year are excluded. 95% confidence intervals are constructed using standard errors clustered at the ORI level.

Figure A8: Impact of budget reform on non-crime economic factors

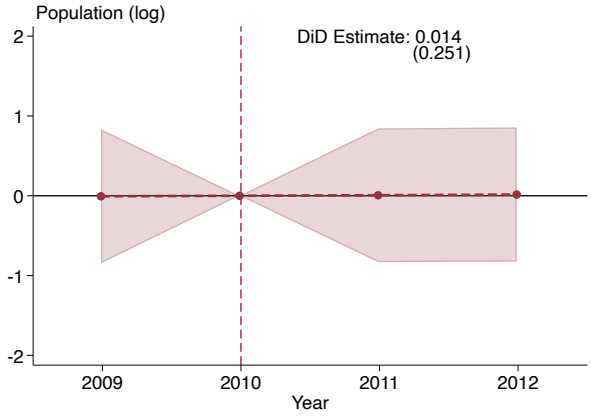
(a) FHFA House Price Index



(b) Unemployment rate



(c) Population (log)



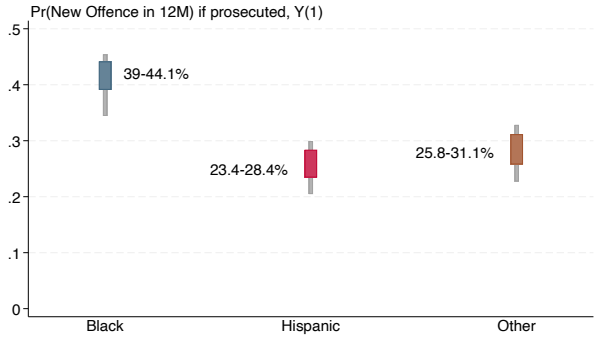
Note: The specification used here is similar to those used to estimate the first stage and reduced form except for the usage of the average pre-reform county-level population as weights in Panels (a) and (b). County-level data on house prices are from the Federal Housing Finance Agency (FHFA), unemployment rates from Local Area Unemployment Statistics and population counts are from the Census Bureau. 95% confidence intervals are constructed using heteroskedasticity-robust standard errors.

Figure A9: Testing for defiers: First stage by subgroup

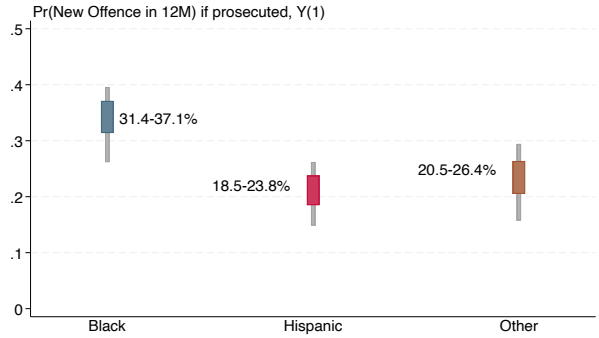
Note: Each Panel presents DiD estimates as reported in Figure 2, but for different covariate subgroups. Age at disposition is split into terciles, represented by T1–T3.

Figure A10: Average re-offence outcomes if prosecuted, disaggregated by minority subgroup

(a) Pre-reform ($Z = 0$)



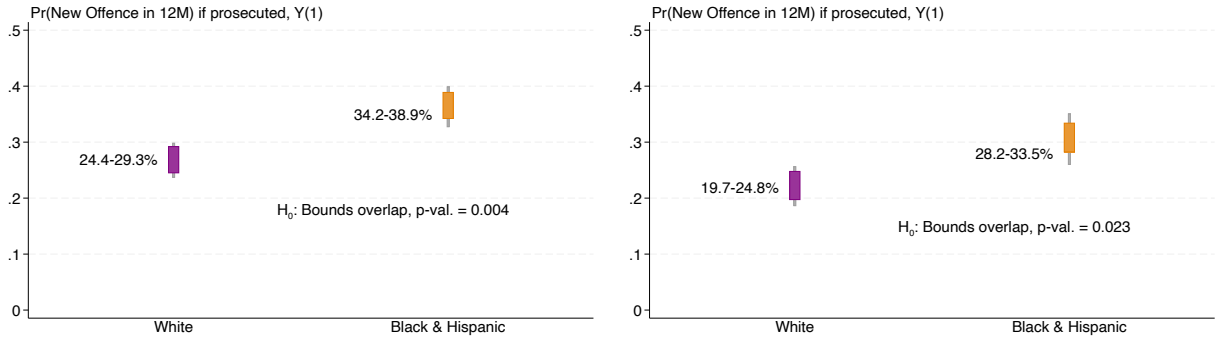
(b) Post-reform ($Z = 1$)



Note: This figure presents bounds on the average treated outcome obtained using the approach described in Section 2.2, separately by time period and subgroups within minority defendants. The treatment is prosecution and the treated outcome, $Y_i(1)$, is whether an individual re-offends one year after disposition, if prosecuted. Confidence intervals are for the true parameter and are bootstrapped using 1,000 replications and a Bayesian bootstrap.

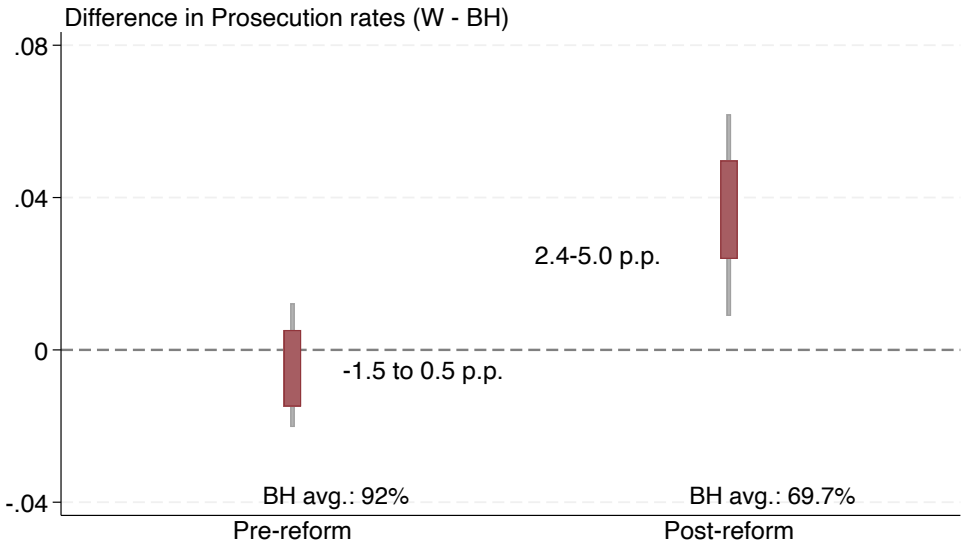
Figure A11: Average re-offence outcomes if prosecuted: White vs. Black/Hispanic defendants

(a) Pre-reform ($Z = 0$) (b) Post-reform ($Z = 1$)



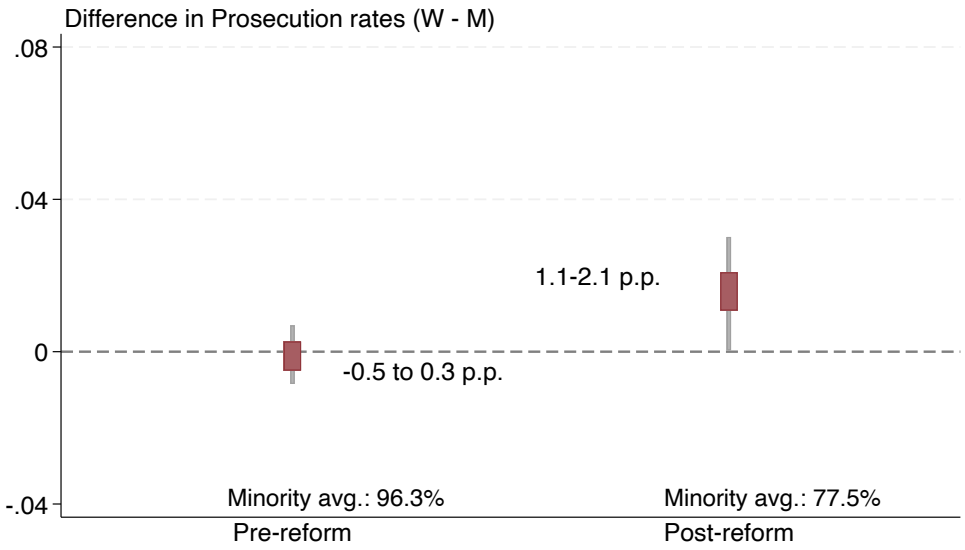
Note: This figure presents bounds on the average treated outcome obtained using the approach described in Section 2.2, separately by race and time period. The treatment is prosecution and the treated outcome, $Y_i(1)$, is whether an individual re-offends one year after disposition, if prosecuted. Confidence intervals are for the true parameter and are bootstrapped using 1,000 replications and a Bayesian bootstrap. The p-value is from a formal bootstrapped test of whether the identified sets overlap, described in Appendix B.6.

Figure A12: Racial prosecution gap conditional on prosecuted outcome: White vs. Black/Hispanic



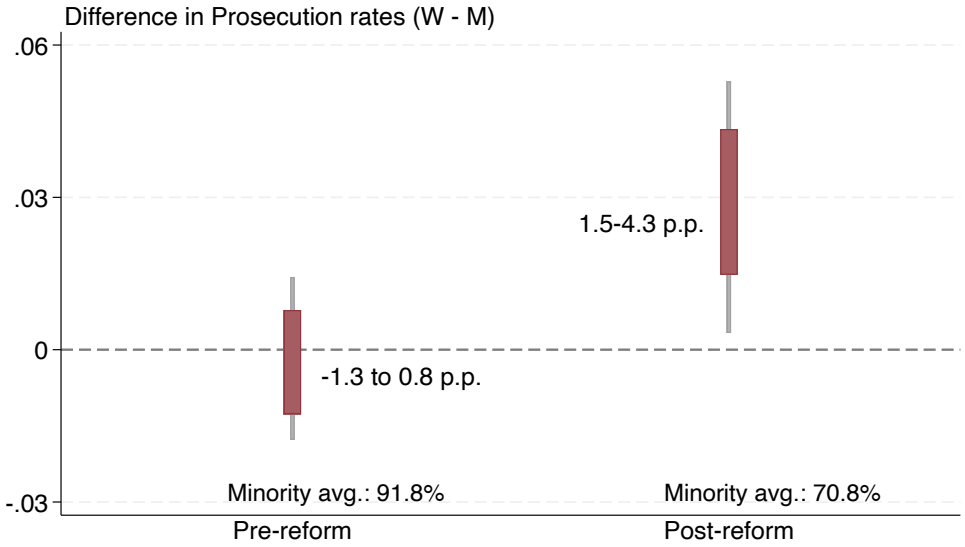
Note: This figure presents bounds on the average difference in prosecution rates in each time period, conditional on prosecuted outcomes, $Y_i(1)$, using the approach described in Section 2.2. $Y_i(1)$ is whether an individual re-offends one year after disposition, if prosecuted. Confidence intervals are for the true parameter and are bootstrapped using 1,000 replications and a Bayesian bootstrap.

Figure A13: Racial prosecution gap cond. on prosecuted outcome: Broad prosecution definition



Note: This figure presents bounds on the average difference in prosecution rates in each time period, conditional on prosecuted outcomes, $Y_i(1)$, using the approach described in Section 2.2. $Y_i(1)$ is whether an individual re-offends one year after disposition, if prosecuted. Confidence intervals are for the true parameter and are bootstrapped using 1,000 replications and a Bayesian bootstrap.

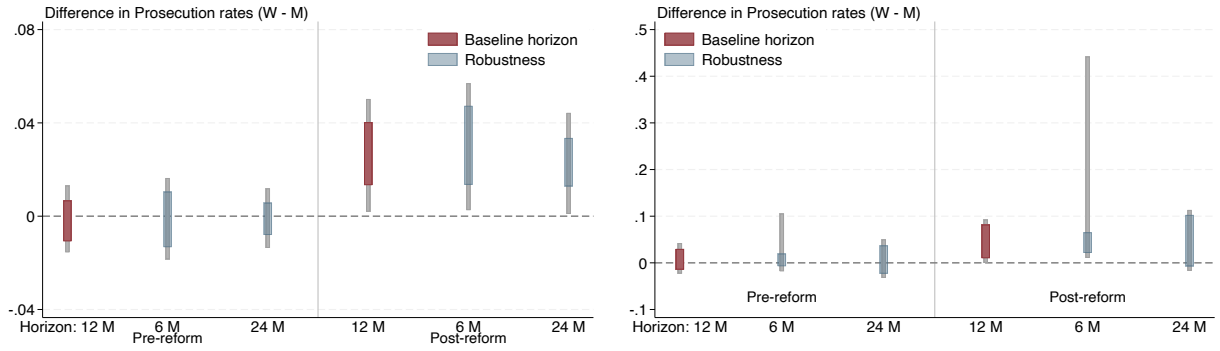
Figure A14: Racial prosecution gap conditional on 'high-priority' re-offence within 1 year



Note: This figure presents bounds on the average difference in prosecution rates in each time period, conditional on prosecuted outcomes, $Y_i(1)$, using the approach described in Section 2.2. $Y_i(1)$ is whether an individual commits a 'high-priority' re-offence one year after disposition if prosecuted, where 'high-priority' offences are those that are not associated with charges that were commonly dismissed in the 2 quarters after the budget reform. Confidence intervals are for the true parameter and are bootstrapped using 1,000 replications and a Bayesian bootstrap.

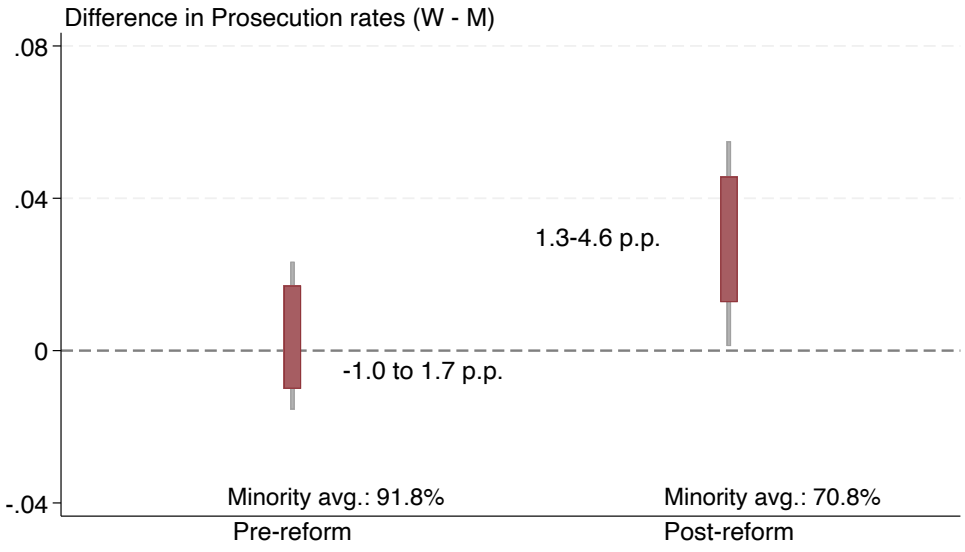
Figure A15: Racial prosecution gap conditional on potential outcomes: Variation by outcome horizon

(a) $Y_i(1)$ (b) $Y_i(0)$



Note: This figure presents bounds on the average difference in prosecution rates in each time period, conditional on potential outcomes, $Y_i(\cdot)$. $Y_i(\cdot)$ is whether an individual re-offends within the amount of time labelled on the x-axis after disposition, if prosecuted (Panel (a)) or if dismissed (Panel (b)). Confidence intervals are for the true parameter and are bootstrapped using 1,000 replications and a Bayesian bootstrap.

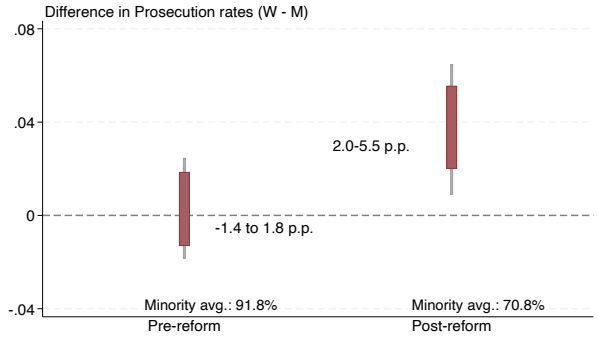
Figure A16: Racial prosecution gap conditional on prosecuted outcome: Worst case bounds



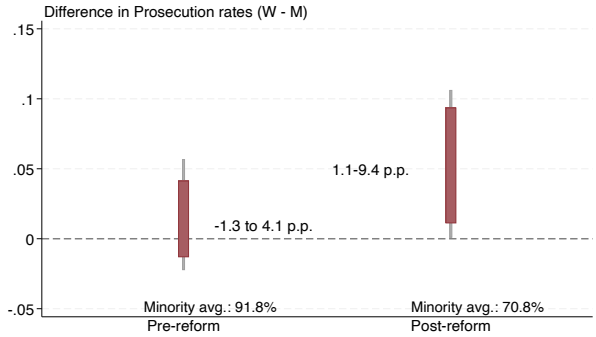
Note: This figure presents bounds on the average difference in prosecution rates in each time period, conditional on prosecuted outcomes, $Y_i(1)$, assuming that never takers' outcomes are bounded between 0 and 1, the widest possible bounds. $Y_i(1)$ is whether an individual re-offends one year after disposition, if prosecuted. Confidence intervals are for the true parameter and are bootstrapped using 1,000 replications and a Bayesian bootstrap.

Figure A17: Racial prosecution gap, conditional on potential outcomes and baseline covariates

(a) $Y_i(1)$



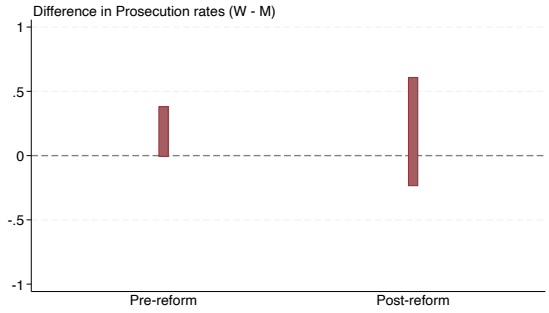
(b) $Y_i(0)$



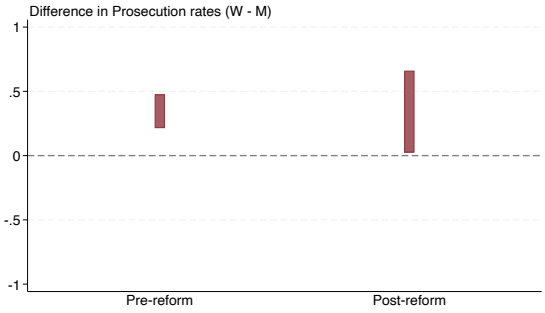
Note: This figure presents bounds on the average difference in prosecution rates in each time period, conditional on potential outcomes, $Y_i(\cdot)$, and baseline covariates. $Y_i(\cdot)$ is whether an individual re-offends one year after disposition if prosecuted (Panel (a)) or if dismissed (Panel (b)). Baseline covariates include age, gender, criminal history (convictions and charges), and court fixed effects. Confidence intervals are for the true parameter and are bootstrapped using 1,000 replications and a Bayesian bootstrap.

Figure A18: Racial prosecution gap, simultaneously conditioning on re-offending ($Y(1)$) and case success ($S(1)$) if prosecuted

(a) $Y(1) = 0$, $S(1) = 0$

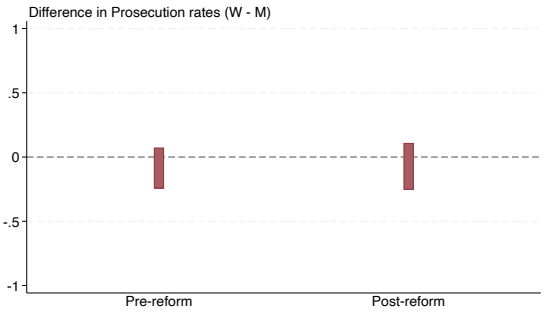


(b) $Y(1) = 1$, $S(1) = 0$



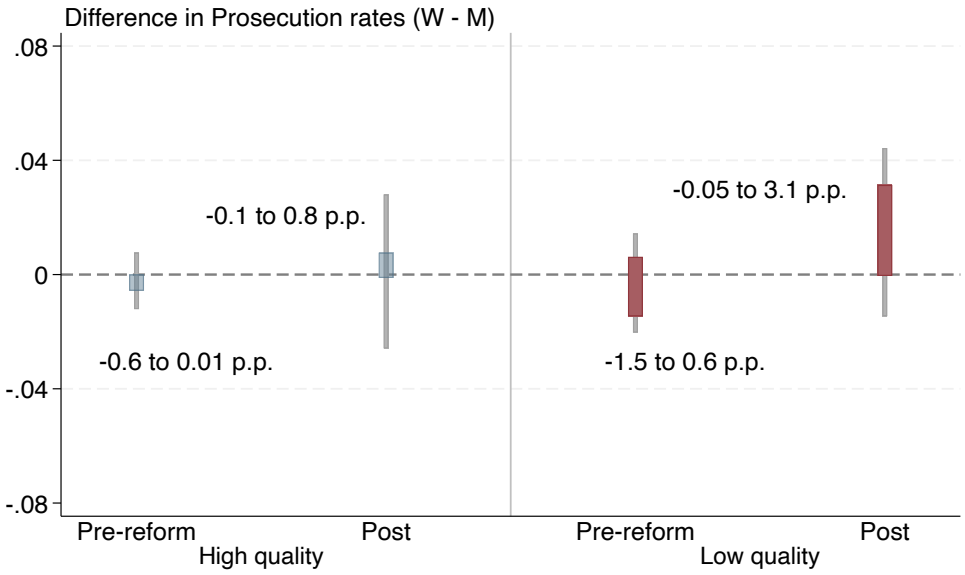
(c) $Y(1) = 0$, $S(1) = 1$

(d) $Y(1) = 1$, $S(1) = 1$



Note: This figure presents bounds on the average difference in prosecution rates in each time period, separately for each combination of the re-offence and case outcome if prosecuted, using the approach described in Appendix B.1. $Y_i(1)$ is whether an individual re-offends one year after disposition if prosecuted and $S(1)$ is whether an individual is sentenced to any non-fine punishment, if prosecuted. Confidence intervals are omitted for readability.

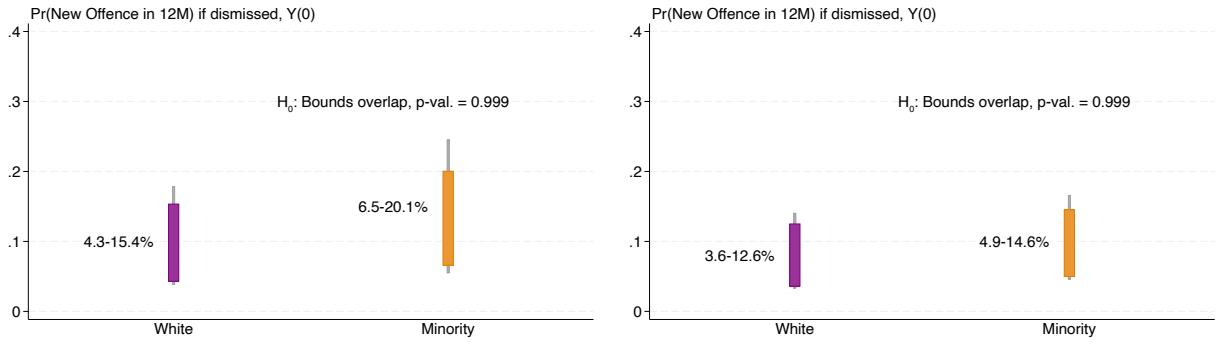
Figure A19: Racial prosecution gap, by proxy for case quality: excluding cases with missing charges



Note: This figure presents bounds on the average difference in prosecution rates in each time period, conditional on prosecuted outcomes, $Y_i(1)$, using the approach described in Section 2.2. $Y_i(1)$ is whether an individual re-offends one year after disposition, if prosecuted. 'High quality'/'Low quality' offences are those with an above/below median share of charges that result in any punishment using pre-reform data, excluding cases where no charges are listed. Confidence intervals are for the true parameter and are bootstrapped using 1,000 replications and a Bayesian bootstrap.

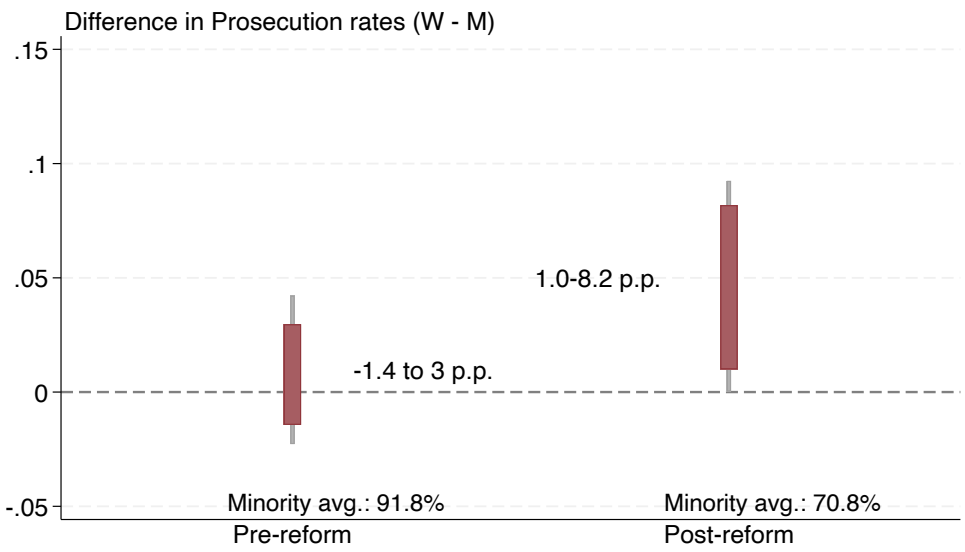
Figure A20: Average re-offence outcomes if dismissed, $E[Y_{it}(0)|R_i = r]$

(a) Pre-reform ($Z = 0$) (b) Post-reform ($Z = 1$)



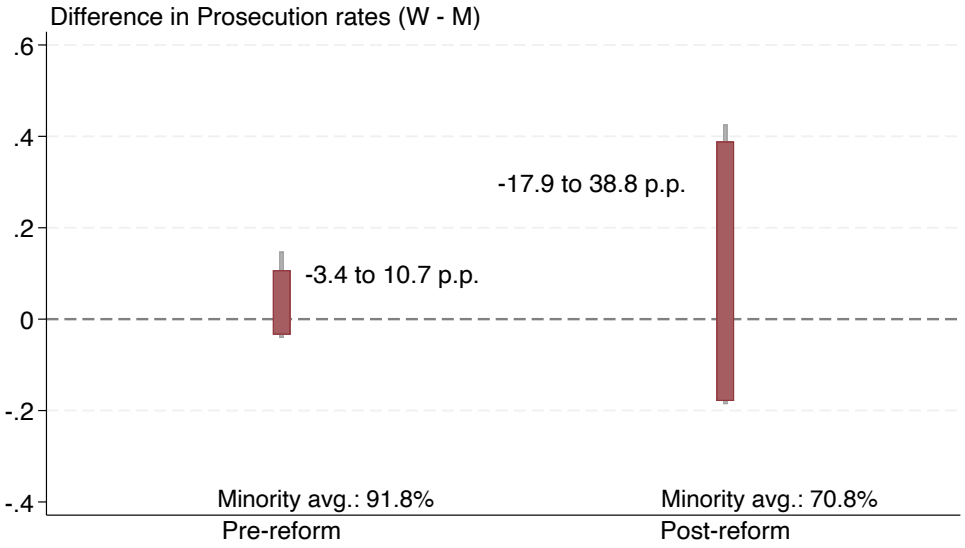
Note: This figure presents bounds on the average untreated outcome obtained using the approach described in Section 2.2, separately by race and time period. The treatment is prosecution and the untreated outcome, $Y_i(0)$, is whether an individual re-offends one year after disposition, if dismissed. Confidence intervals are for the true parameter and are bootstrapped using 1,000 replications and a Bayesian bootstrap.

Figure A21: Racial prosecution gap conditional on dismissed outcome



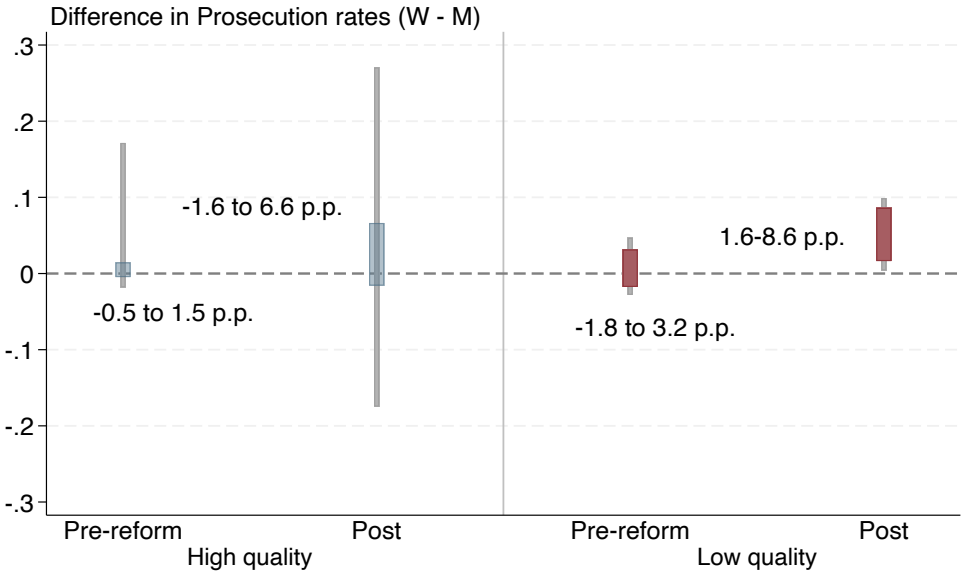
Note: This figure presents bounds on the average difference in prosecution rates in each time period, conditional on dismissed outcomes, $Y_i(0)$, using the approach described in Section 2.2. $Y_i(0)$ is whether an individual re-offends one year after disposition, if dismissed. Confidence intervals are for the true parameter and are bootstrapped using 1,000 replications and a Bayesian bootstrap.

Figure A22: Racial prosecution gap conditional on dismissed outcome: Worst case bounds



Note: This figure presents bounds on the average difference in prosecution rates in each time period, conditional on dismissed outcomes, $Y_i(0)$, , assuming that always takers' outcomes are bounded between 0 and 1, the widest possible bounds. $Y_i(0)$ is whether an individual re-offends one year after disposition, if dismissed. Confidence intervals are for the true parameter and are bootstrapped using 1,000 replications and a Bayesian bootstrap.

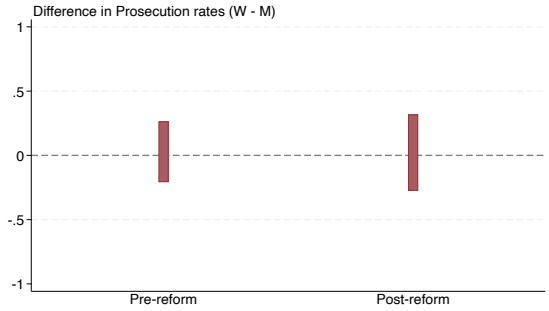
Figure A23: Racial prosecution gap, by proxy for case quality: conditional on dismissed outcome



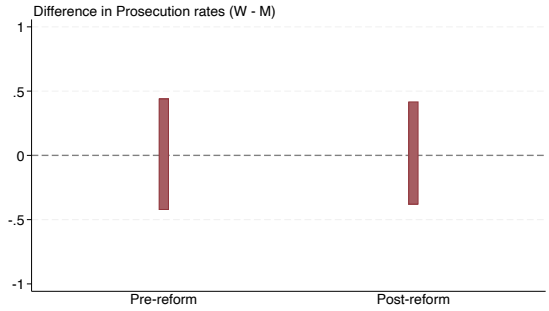
Note: This figure presents bounds on the average difference in prosecution rates in each time period, conditional on dismissed outcomes, $Y_i(0)$, using the approach described in Section 2.2. $Y_i(0)$ is whether an individual re-offends one year after disposition, if dismissed. 'High quality'/'Low quality' offences are those with an above/below median share of charges that result in any punishment using pre-reform data. Confidence intervals are for the true parameter and are bootstrapped using 1,000 replications and a Bayesian bootstrap.

Figure A24: Racial prosecution gap, simultaneously conditioning on outcome if prosecuted ($Y(1)$) and dismissed ($Y(0)$)

(a) $Y(1) = 0$, $Y(0) = 0$



(b) $Y(1) = 1$, $Y(0) = 0$



(c) $Y(1) = 0$, $Y(0) = 1$

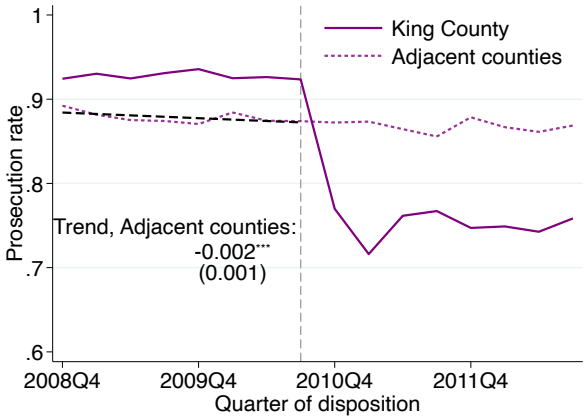
(d) $Y(1) = 1$, $Y(0) = 1$

Note: This figure presents bounds on the average difference in prosecution rates in each time period, separately for each combination of outcomes if prosecuted and dismissed, using the approach described in Appendix B.1. $Y_i(1)$ is whether an individual re-offends one year after disposition if prosecuted and $Y_i(0)$ is whether an individual re-offends one year after disposition if dismissed. Confidence intervals are omitted for readability.

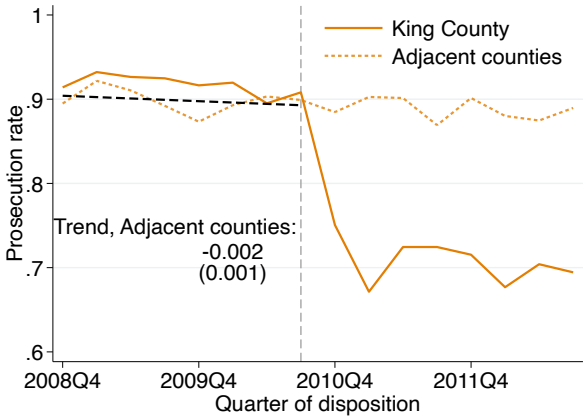
A.2 Empirically validating DiD adjustment assumptions

Figure A25: Testing trends in prosecution rates in adjacent counties

(a) White defendants



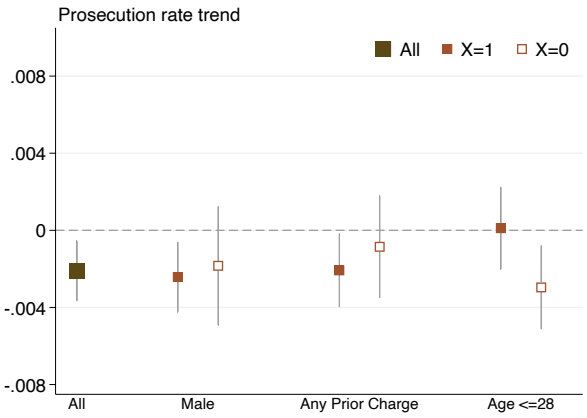
(b) Minority defendants



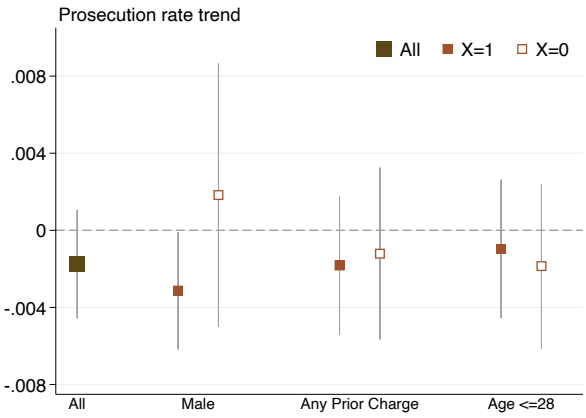
Note: The displayed coefficients are from estimating a linear regression of prosecution on a linear trend using pre-period data in the counties adjacent to King County. Standard errors on pre-period trend estimates are heteroscedasticity-robust.

Figure A26: Prosecution trends in adjacent counties, by covariate subgroup

(a) White defendants



(b) Minority defendants



Note: Each coefficient is from estimating a linear regression of prosecution on a linear quarterly trend using pre-period data on individuals from a given subgroup (denoted by X) in the counties adjacent to King County. The estimate labelled 'All' reproduces the overall trend estimate from Figure A25. 95% confidence intervals are constructed using heteroscedasticity-robust standard errors.

Figures A27–A30 estimate the following regression; Figure A27 and Figure A29 test hypotheses 1) and 2), while Figure A28 and Figure A30 test hypotheses 3) and 4):

$$
\begin{aligned}
Y_{itg} = {} & \beta_1 \mathrm{t} + \beta_2 X_i + \beta_3 \text{King County} + \\
& \delta_1 X_i \times \mathrm{t} + \delta_2 X_i \times \text{King County} + \delta_3 \mathrm{t} \times \text{King County} + \\
& \delta_4 X_i \times \mathrm{t} \times \text{King County} + \varepsilon_{igt}
\end{aligned}
$$

1) $X = 0$, $H_0 : \delta_3 = 0$

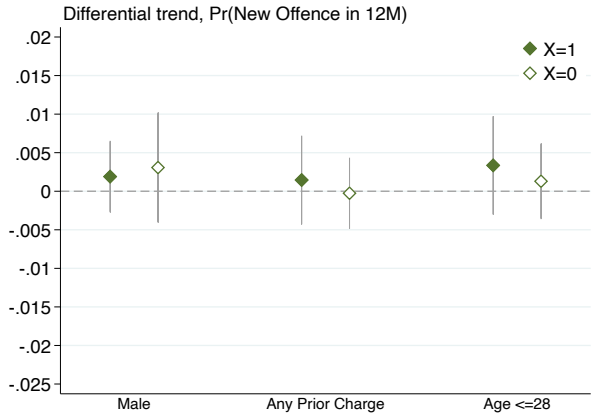
2) $X = 1$, $H_0 : \delta_3 + \delta_4 = 0$

3) King County, $H_0 : \delta_1 + \delta_4 = 0$

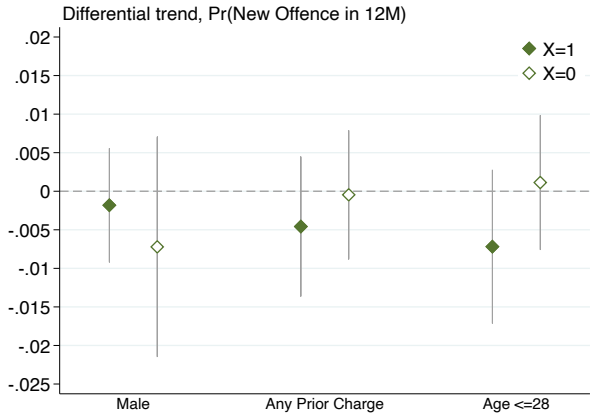
4) Adjacent, $H_0 : \delta_1 = 0$

Figure A27: Testing for differential trends in group-specific treated outcomes across counties

(a) White defendants



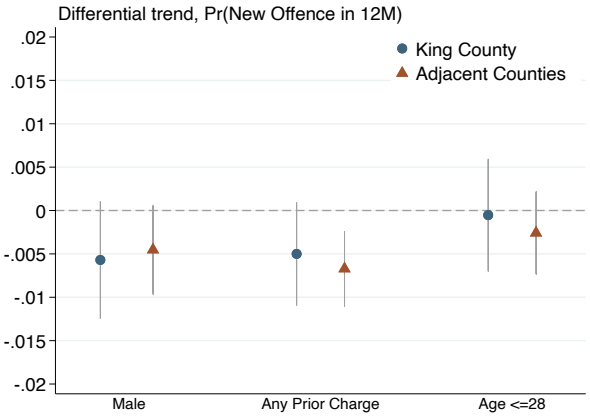
(b) Minority defendants



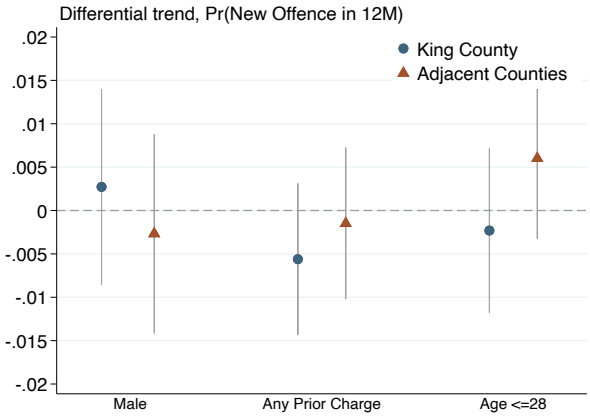
Note: Each coefficient is an estimate of the difference in pre-period trends in treated outcomes (outcomes if prosecuted) across counties, for a given covariate value. The sample only includes individuals in our baseline analysis sample who are prosecuted prior to the budget reform. For example, the first solid green diamond in Panel a) represents the difference in trends in treated outcomes for white male defendants between King County and other counties. 95% confidence intervals are constructed using heteroscedasticity-robust standard errors.

Figure A28: Testing for differential trends in treated outcomes between groups, within each county

(a) White defendants



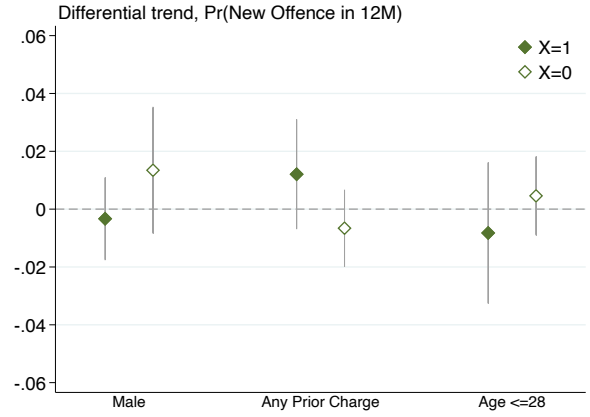
(b) Minority defendants



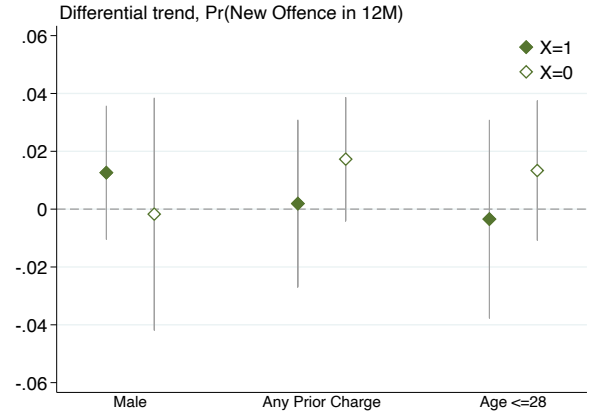
Note: Each coefficient is an estimate of the difference in pre-period trends in treated outcomes (outcomes if prosecuted) across covariate groups, within a given county. The sample only includes individuals in our baseline analysis sample who are prosecuted prior to the budget reform. For example, the first blue circle in Panel a) represents the difference in trends in treated outcomes between white male and female defendants in King County. 95% confidence intervals are constructed using heteroscedasticity-robust standard errors.

Figure A29: Testing for differential trends in group-specific untreated outcomes across counties

(a) White defendants



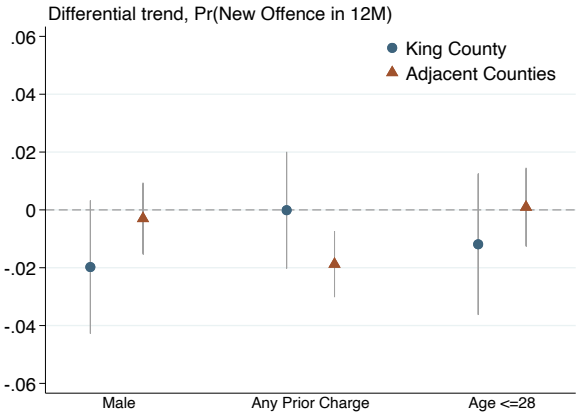
(b) Minority defendants



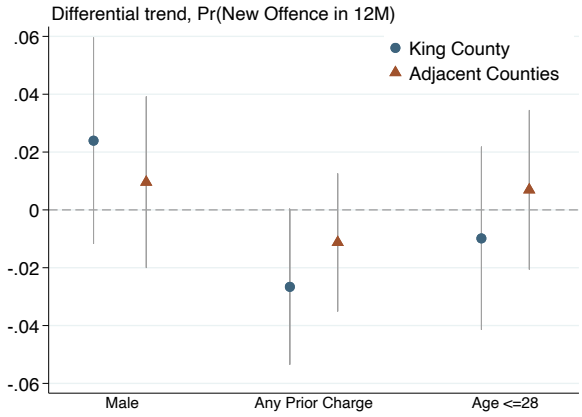
Note: Each coefficient is an estimate of the difference in pre-period trends in untreated outcomes (outcomes if not prosecuted) across counties, for a given covariate value. The sample only includes individuals in our baseline analysis sample who are dismissed prior to the budget reform. For example, the first green diamond in Panel a) represents the difference in trends in untreated outcomes for white male defendants between King County and other counties. 95% confidence intervals are constructed using heteroscedasticity-robust standard errors.

Figure A30: Testing for differential trends in untreated outcomes between groups, within each county

(a) White defendants



(b) Minority defendants



Note: Each coefficient is an estimate of the difference in pre-period trends in untreated outcomes (outcomes if not prosecuted) across covariate groups, within a given County. The sample only includes individuals in our baseline analysis sample who are dismissed prior to the budget reform. For example, the first blue circle in Panel a) represents the difference in trends in untreated outcomes between white male and female defendants in King County. 95% confidence intervals are constructed using heteroscedasticity-robust standard errors.

Appendix B Methodological Details

B.1 Discrimination conditional on other functions of potential outcomes

We have discussed estimating discrimination conditional on **treated** potential outcomes. However, it might be more appropriate in certain contexts to measure discrimination as differential treatment among individuals with the same **untreated** potential outcomes, the same **treatment effect**, or even **multiple** distinct treated/untreated potential outcomes. Each of these may also map to different normative notions of fairness. We next discuss estimating these other discrimination estimands, again assuming that potential outcomes are binary for simplicity.

Conditioning on untreated potential outcomes, $Y_i(0)$

Differential treatment among individuals with the same untreated potential outcome is a function of period-, race-, and potential outcome-specific treatment rates, as shown in Equation 5. This diverges from the treatment rates conditional on treated potential outcomes from Equation 3 in two ways. 1) The denominator is now the average outcome that would be realized if **no one was treated**. 2) The first term in the numerator, the average **untreated** outcome among those who were **treated**, is no longer directly observed in the data, since always takers are always treated.

$$E[D_i|Z=z,R_i=r,Y_i(0)=1] = \frac{E[Y_i(0)=1|Z=z,R_i=r,D_i=1]\times E[D_i|Z=z,R_i=r]}{E[Y_i(0)=1|R_i=r]} \quad (5)$$

Assumptions on the relationship between treatment propensity and average **untreated** potential outcomes, analogous to those described for treated outcomes, deal with both instances of divergence. These assumptions bound or point-identify the average outcomes if no one were treated, $E[Y_i(0)=y|R_i=r]$. Extrapolating $E[Y_i(0)=y|R_i=r]$ involves extrapolating the average untreated outcomes of always takers (since they are always treated) using estimates of the untreated outcomes of compliers and never takers. Note that always takers' untreated outcomes are a component of $E[Y_i(0)=y|Z=z,R_i=r,D_i=1]$. Plugging in bounds/point estimates from each of these steps into Equation 5 yields bounds/point estimates for $E[D_i|Z=z,R_i=r,Y_i(0)=y]$. These treatment rates would then be aggregated up, in a way analogous to Equation 3, to estimate group differences in treatment among those who would have identical outcomes if not treated.

Alternatively, since treatment is binary here, the **prosecution rate** conditional on outcome if dismissed is equivalent to subtracting the **dismissal rate** for those who would have that outcome if dismissed from 1. To see this, consider Equation 6, which shows how the share of individuals **not treated**, with a given untreated outcome (e.g., the dismissal rate for people with a specific outcome if dismissed) can be used to quantify the share of **treated** individuals with a specific untreated outcome, abstracting away from the period-specific notation Z. The first line subtracts the share of individuals not treated with a given untreated outcome from 1, where $\bar{D}=E[D_i|R_i=r]$, and where $E[D_i=0|R_i=r,Y_i(0)=y]$ is rewritten in the same way as described in Equation 1. The

third line uses the fact that the average untreated outcome observed if everyone were untreated is a weighted average of the untreated outcomes for the treated individuals and the untreated individuals: $E[Y_i(0) = y|R_i = r] = E[Y_i(0) = y|R_i = r, D_i = 1]\bar{D} + E[Y_i(0) = y|R_i = r, D_i = 0](1-\bar{D})$. This recovers the share of treated individuals with a specific untreated outcome, described in Equation 5.

$$
\begin{aligned}
1 - E[D_i = 0|R_i = r, Y_i(0) = y] &= 1 - \frac{E[Y_i(0) = y|R_i = r, D_i = 0] \times (1 - \bar{D})}{E[Y_i(0) = y|R_i = r]} \\
&= \frac{E[Y_i(0) = y|R_i = r] - E[Y_i(0) = y|R_i = r, D_i = 0] \times (1 - \bar{D})}{E[Y_i(0) = y|R_i = r]} \\
&= \frac{E[Y_i(0) = y|R_i = r, D_i = 1] \times \bar{D}}{E[Y_i(0) = y|R_i = r]} \\
&= E[D_i|R_i = r, Y_i(0) = y]
\end{aligned} \tag{6}
$$

Given this mapping, if one was interested in estimating discrimination conditional on untreated outcomes, an attractive natural experiment is one that generates a small share of always takers. While this discussion has focused on a simple binary IV, Appendix B.5 discusses the assumptions required to estimate discrimination conditional on untreated potential outcomes with DiD variation.

Conditioning on the joint distribution of potential outcomes $(Y_i(0), Y_i(1))$ or treatment effects $(Y_i(1) - Y_i(0))$

In the main text, we show how to estimate discrimination using a binary instrumental variable, conditional on a single binary potential outcome, $Y_i(1)$. However, researchers may be interested in quantifying discrimination conditional on the *joint distribution* of potential outcomes $(Y_i(0), Y_i(1))$ or *treatment effects* $(Y_i(1) - Y_i(0))$. Conditioning on either of these quantities is difficult when $Y_i(0)$ is not mechanically zero, since conditioning on $Y_i(1)$ in that special case is equivalent to conditioning on the treatment effect (e.g., in the bail context). More generally, researchers may also wish to condition on the joint distribution of *multiple treated* (or untreated) potential outcomes (e.g., recidivism and conviction success). The method we outline here permits this type of conditioning as well.

In this appendix, we illustrate how to estimate discrimination conditional on the joint distribution of potential outcomes, treatment effects, or multiple treated (or untreated) potential outcomes. We begin by showing how to identify key moments—the average joint potential outcomes. We then highlight their connection to conditioning on treatment effects, which is nested in the general joint potential outcome case. We also discuss how to use information from the marginal distributions to tighten the bounds on key objects when simultaneously conditioning on multiple potential outcomes. For notational simplicity, we assume that average potential outcomes $E[Y_i(1)|\cdot], E[Y_i(0)|\cdot]$ are point identified, although the method outline below also goes through with partially identified

objects as well. For simplicity, we abstract from any $Z_i \in \{0,1\}$ dimension in the notation below.

Conditioning on the joint distribution of potential outcomes $(Y_i(0), Y_i(1))$
Expanding the approach from the main text to simultaneously condition on pairs of potential outcomes is not straightforward for the following reasons: 1) the covariance between potential outcomes is fundamentally unobserved, 2) the joint distribution of potential outcomes is not point identified even if the marginal distributions, e.g., $E[Y_i(1) = 1]$, are point identified.

Setup: With simultaneously conditioning on pairs of potential outcomes, define the conditional treatment rates, $\pi_{r(y_0,y_1)}$, as:

$$\pi_{r(y_0,y_1)} = E[D_i | R_i = r, Y_i(0) = y_0, Y_i(1) = y_1]$$

for $r \in \{r_1, r_2\}$ and binary $Y_i(0), Y_i(1)$. Using the entire distribution of potential outcomes and similar arguments to the unidimensional case in the main text, we can re-write each $\pi_{r(y_0,y_1)}$:

$$\begin{aligned}
\pi_{r(0,0)} &= \frac{E[Y_i(0) = 0, Y_i(1) = 0 | R_i = r, D_i = 1] \times E[D_i | R_i = r]}{E[Y_i(0) = 0, Y_i(1) = 0 | R_i = r]} \\
\pi_{r(1,0)} &= \frac{E[Y_i(0) = 1, Y_i(1) = 0 | R_i = r, D_i = 1] \times E[D_i | R_i = r]}{E[Y_i(0) = 1, Y_i(1) = 0 | R_i = r]} \\
\pi_{r(0,1)} &= \frac{E[Y_i(0) = 0, Y_i(1) = 1 | R_i = r, D_i = 1] \times E[D_i | R_i = r]}{E[Y_i(0) = 0, Y_i(1) = 1 | R_i = r]} \\
\pi_{r(1,1)} &= \frac{E[Y_i(0) = 1, Y_i(1) = 1 | R_i = r, D_i = 1] \times E[D_i | R_i = r]}{E[Y_i(0) = 1, Y_i(1) = 1 | R_i = r]}
\end{aligned} \tag{7}$$

Numerator of $\pi_{r(y_0,y_1)}$*:* The key empirical challenge is that only the second term in the numerator is identified in the data. The first term is not observed since it is a function of treated *and* untreated potential outcomes, *conditional on being treated.* Despite being unobserved, we can bound this term by re-writing it as a function of known and unknown quantities using the definition of expected value for two dependent variables.[43] Taking (y_0, y_1) as (1,1) for illustrative purposes, we can then re-write the unknown component of the numerator of $\pi_{r(1,1)}$ as:

$$\begin{aligned}
E[Y_i(0) = 1, Y_i(1) = 1 | R_i = r, D_i = 1] = &\overbrace{E[Y_i(0) = 1 | R_i = r, D_i = 1]}^{\text{Extrapolated}} \times \overbrace{E[Y_i(1) = 1 | R_i = r, D_i = 1]}^{\text{Observed in data}} \\
&+ \underbrace{Cov(Y_i(0), Y_i(1) | R_i = r, D_i = 1)}_{\text{To be bounded}}
\end{aligned} \tag{8}$$

This re-writing reveals what we do and do not need to estimate. $E[Y_i(0) = 1 | R_i = r, D_i = 1]$ (the *untreated* potential outcome, *conditional on being treated*) is identified following the extrapolation

[43] Specifically, this re-writing uses the identity of $E[XY] = E[X]E[Y] + Cov(X, Y)$.

procedure outlined in the previous appendix subsection and $E[Y_i(1) = 1|R_i = r, D_i = 1]$ is identified in the data, as before in the main text. The remaining task is to determine the covariance between potential outcomes, which is fundamentally unknown. Based on the conditional expectations that we have identified, we can generically bound the covariance between $Y_i(0) \in [a, b]$ and $Y_i(1) \in [c, d]$, following Hössjer and Sjölander (2022), suppressing R_i, D_i notation in the expectations for compactness $(\cdot)$, as:

$$\begin{aligned}
&-\min\Big[\big(E[Y_i(0)=1|\cdot]-a\big)\big(E[Y_i(1)=1|\cdot]-c\big), \big(b-E[Y_i(0)=1|\cdot]\big)\big(d-E[Y_i(1)=1|\cdot]\big)\Big] \\
&\leq Cov(Y_i(0), Y_i(1)|R_i = r, D_i = 1) \\
&\leq \min\Big[\big(E[Y_i(0)=1|\cdot]-a\big)\big(d-E[Y_i(1)=1|\cdot]\big), \big(b-E[Y_i(0)=1|\cdot]\big)\big(E[Y_i(1)=1|\cdot]-c\big)\Big]
\end{aligned} \tag{9}$$

which, after incorporating the fact that binary potential outcomes must lie in $[0, 1]$, gives the following covariance bounds:

$$\begin{aligned}
&-\min\Big[\big(E[Y_i(0)=1|\cdot]\big)\big(E[Y_i(1)=1|\cdot]\big), \big(1-E[Y_i(0)=1|\cdot]\big)\big(1-E[Y_i(1)=1|\cdot]\big)\Big] \\
&\leq Cov(Y_i(0), Y_i(1)|R_i = r, D_i = 1) \\
&\leq \min\Big[\big(E[Y_i(0)=1|\cdot]\big)\big(1-E[Y_i(1)=1|\cdot]\big), \big(1-E[Y_i(0)=1|\cdot]\big)\big(E[Y_i(1)=1|\cdot]\big)\Big]
\end{aligned}$$

which we denote as $[\underline{C}_{rd(y_0,y_1)}, \overline{C}_{rd(y_0,y_1)}]$. Plugging in these covariance bounds into Equation 8 yields the following lower ($\underline{E}$) and upper ($\overline{E}$) bounds on the unknown object in the numerator of $\pi_{r(1,1)}$:

$$\begin{aligned}
\underline{E}[Y_i(0)=1, Y_i(1)=1|R_i=r, D_i=1] =& \overbrace{E[Y_i(0)=1|R_i=r, D_i=1]}^{\text{Extrapolated}} \times \overbrace{E[Y_i(1)=1|R_i=r, D_i=1]}^{\text{Observed in data}} \\
&+ \underbrace{\underline{C}_{r1(1,1)}}_{\substack{\text{Lower} \\ \text{covariance bound}}} \\
\overline{E}[Y_i(0)=1, Y_i(1)=1|R_i=r, D_i=1] =& \overbrace{E[Y_i(0)=1|R_i=r, D_i=1]}^{\text{Extrapolated}} \times \overbrace{E[Y_i(1)=1|R_i=r, D_i=1]}^{\text{Observed in data}} \\
&+ \underbrace{\overline{C}_{r1(1,1)}}_{\substack{\text{Upper} \\ \text{covariance bound}}}
\end{aligned}$$

While we have illustrated this bounding approach using $Y_i(0) = Y_i(1) = 1$ as the leading example, it is straightforward to apply this method to bound other combinations of $(Y_i(0) = y_0, Y_i(1) = y_1)$ by redefining potential outcomes using combinations of the appropriate indicator variables $(\mathbb{I}[Y_i(0) = 0], \mathbb{I}[Y_i(0) = 1], \mathbb{I}[Y_i(1) = 0], \mathbb{I}[Y_i(1) = 1])$. For example, bounding $E[Y_i(0) = 0, Y_i(1) = 1|R_i = r, D_i = 1]$ would define the focal potential outcomes using $\mathbb{I}[Y_i(0) = 0]$ and $\mathbb{I}[Y_i(1) = 1]$

indicators. Note that the bounds on the covariance of $(Y_i(0), Y_i(1))$ will depend on the indicator functions used. For binary potential outcomes, the covariance bounds will coincide with each other (and with $Cov(Y_i(0), Y_i(1)|R_i = r, D_i = 1)$) when $y_0 = y_1$ and will be their negation when $y_0 \neq y_1$.[44]

Denominator of $\pi_{r(y_0,y_1)}$: We now turn our focus to the denominator. Here, we are not interested in the treated conditional expectation, but the overall $r \in R_i$ sub-population expectation, *unconditional on treatment status.* We can follow a similar procedure as outlined for the numerator above to obtain bounds on the sub-population joint distribution of potential outcomes, recalling that for notational simplicity we assume that the marginal distributions are point identified, and taking (y_0, y_1) as (1,1):

$$\begin{aligned} E[Y_i(0) = 1, Y_i(1) = 1|R_i = r] = &\overbrace{E[Y_i(0) = 1|R_i = r]}^{\text{Extrapolated}} \times \overbrace{E[Y_i(1) = 1|R_i = r]}^{\text{Extrapolated}} \\ &+ \underbrace{Cov(Y_i(0, Y_i(1)|R_i = r)}_{\text{To be bounded}} \end{aligned} \tag{10}$$

The bounds on the covariance are given by the following expression, again using the fact that potential outcomes are binary:

$$\begin{aligned} &-\min\Big[\big(E[Y_i(0) = 1|R_i = r]\big)\big(E[Y_i(1) = 1|R_i = r]\big), \big(1 - E[Y_i(0) = 1|R_i = r]\big)\big(1 - E[Y_i(1) = 1|R_i = r]\big)\Big] \\ \leq &Cov(Y_i(0), Y_i(1)|R_i = r) \\ \leq &\min\Big[\big(E[Y_i(0) = 1|R_i = r]\big)\big(1 - E[Y_i(1) = 1|R_i = r]\big), \big(1 - E[Y_i(0) = 1|R_i = r]\big)\big(E[Y_i(1) = 1|R_i = r]\big)\Big] \end{aligned}$$

We denote the bounds on the population covariance (which may be different than the treated subsample covariance) as $\underline{C}_{r(y_0,y_1)}$, $\overline{C}_{r(y_0,y_1)}$. As before, we can insert these covariance bounds into the unknown joint expectations of interest to obtain bounds on the group-specific average outcomes:

$$\begin{aligned} \underline{E}[Y_i(0) = 1, Y_i(1) = 1|R_i = r] = &\overbrace{E[Y_i(0) = 1|R_i = r]}^{\text{Extrapolated}} \times \overbrace{E[Y_i(1) = 1|R_i = r]}^{\text{Extrapolated}} \\ &+ \quad \underbrace{\underline{C}_{r(1,1)}}_{\substack{\text{Lower} \\ \text{covariance bound}}} \\ \overline{E}[Y_i(0) = 1, Y_i(1) = 1|R_i = r] = &\overbrace{E[Y_i(0) = 1|R_i = r]}^{\text{Extrapolated}} \times \overbrace{E[Y_i(1) = 1|R_i = r]}^{\text{Extrapolated}} \\ &+ \quad \underbrace{\overline{C}_{r(1,1)}}_{\substack{\text{Upper} \\ \text{covariance bound}}} \end{aligned}$$

[44] Cauchy-Schwarz bounds provide a feasible alternative (for binary potential outcomes) and will be invariant to the set of indicator functions used. However, they may not be as sharp as the Hössjer and Sjölander (2022) bounds in general.

These bounds on group-specific outcomes can be combined with the bounds on group-specific *treated* outcomes derived in Equation 7 to obtain bounds on $\pi_{r(y_0,y_1)}$ for all (y_0, y_1) combinations, which we denote as $[\underline{\pi}_{r(y_0,y_1)}, \overline{\pi}_{r(y_0,y_1)}]$.

Potential outcome shares: The remaining item to calculate to go from the conditional treatment rates, π, to the discrimination quantity, $\triangle$ (Equation 3), are the population shares in each cell of the discrete potential outcome distribution, $Pr(Y_i(0) = y_0, Y_i(1) = y_1)$. In the unidimensional case, calculating these is trivial since $Pr(Y_i(1) = 1)$ and $Pr(Y_i(1) = 0) = 1 - Pr(Y_i(1) = 1)$. In the case of binary potential outcomes, these shares are already bounded for each racial subgroup when bounding the denominator of the $\pi_{r(y_0,y_1)}$ terms as $\underline{E}[Y_i(0) = y_0, Y_i(1) = y_1|R_i = r]$ and $\overline{E}[Y_i(0) = y_0, Y_i(1) = y_1|R_i = r]$. Specifically, bounds on the group-specific potential outcome shares are given by the following:

$$[\underline{p}_{r(y_0,y_1)}, \overline{p}_{r(y_0,y_1)}] = \left[\underline{E}[Y_i(0) = 1, Y_i(1) = 1|R_i = r], \overline{E}[Y_i(0) = 1, Y_i(1) = 1|R_i = r]\right] \tag{11}$$

$$= \left[\underline{Pr}(Y_i(0) = 1, Y_i(1) = 1|R_i = r), \overline{Pr}(Y_i(0) = 1, Y_i(1) = 1|R_i = r)\right] \tag{12}$$

where the expectations on the right-hand side of equation (11) are derived above and going from expectations to probabilities is by virtue of binary potential outcomes. Finally, since $Pr(R_i = r)$ is known in the data, it is straightforward to construct bounds on the population-level shares using the derived bounds and each group r's prevalence.

$$[\underline{p}_{(y_0,y_1)}, \overline{p}_{(y_0,y_1)}] = [\underline{p}_{r_1(y_0,y_1)} Pr(R_i = r_1) + \underline{p}_{r_2(y_0,y_1)}(1 - Pr(R_i = r_1)),$$
$$\overline{p}_{r_1(y_0,y_1)} Pr(R_i = r_1) + \overline{p}_{r_2(y_0,y_1)}(1 - Pr(R_i = r_1))]$$

Quantifying discrimination: We now have all of the required pieces to estimate bounds on discrimination, conditional on the entire joint distribution of treated and untreated potential outcomes.[45]

$$\triangle_{(y_0,y_1)} = \pi_{r_1(y_0,y_1)} - \pi_{r_2(y_0,y_1)} \tag{13}$$

$$\triangle = \sum_{y_0 \in supp(Y_i(0))} \sum_{y_1 \in supp(Y_i(1))} p_{(y_0,y_1)} \triangle_{(y_0,y_1)} \tag{14}$$

As described in the main text, we would then plug in the appropriate bounds for each $\pi_{r(y_0,y_1)}$ and $p_{(y_0,y_1)}$ object and then conduct a grid search to find the minimum and maximum values.

Connection to conditioning on treatment effects, $Y_i(1) - Y_i(0)$

Conditioning on the treatment effect, $\tau_i = Y_i(1) - Y_i(0)$, is a natural consequence of conditioning on the entire joint distribution of treated and untreated potential outcomes. Each cell in the potential

[45] For notational simplicity we express $\triangle$ as being a function of point identified objects, following the notation in the main text.

outcome distribution represents a specific value of the treatment effect. For example, with binary potential outcomes, the potential values of the treatment effect are given in the following table:

Table B1: Treatment Effects and Potential Outcomes

Potential Outcomes	Treatment Effect (τ_i)	Treatment Effect Interpretation
$Y_i(1) = 1, Y_i(0) = 0$	$Y_i(1) - Y_i(0) = 1$	Treatment has positive effect
$Y_i(1) = 1, Y_i(0) = 1$	$Y_i(1) - Y_i(0) = 0$	Treatment has no effect
$Y_i(1) = 0, Y_i(0) = 0$	$Y_i(1) - Y_i(0) = 0$	Treatment has no effect
$Y_i(1) = 0, Y_i(0) = 1$	$Y_i(1) - Y_i(0) = -1$	Treatment has negative effect

It is straightforward to see how $\triangle$ averages across not only the joint distribution of treated and untreated potential outcomes, but by consequence, the joint distribution of treatment effects:

$$\triangle = \underbrace{Pr(Y_i(0) = 0, Y_i(1) = 0)\triangle_{(0,0)}}_{\text{Treatment has no effect: } \tau_i = 0} + \underbrace{Pr(Y_i(0) = 1, Y_i(1) = 0)\triangle_{(1,0)}}_{\text{Treatment has negative effect: } \tau_i = -1}$$
$$+ \underbrace{Pr(Y_i(0) = 0, Y_i(1) = 1)\triangle_{(0,1)}}_{\text{Treatment has positive effect: } \tau_i = 1} + \underbrace{Pr(Y_i(0) = 1, Y_i(1) = 1)\triangle_{(1,1)}}_{\text{Treatment has no effect: } \tau_i = 0}$$

Thus, conditioning simultaneously on the treated and untreated potential outcome is equivalent to quantifying discrimination among subsets of the population who have the same treatment effect. Examining each of these components may be useful to understand where in the potential outcome (and treatment effect) distribution disparities are concentrated.

Remark 1 (Tightening bounds through logical consistency): Information from the marginal distributions of $Pr[Y_i(0) = y_0|R_i = r]$ and $Pr[Y_i(1) = y_1|R_i = r]$ can potentially be used to restrict the set of admissible population shares and therefore, the covariances and average joint potential outcomes that are consistent with the data. To see this, note that the following relationships must hold, defining $Pr(Y_i(0) = y_0, Y_i(1) = y_1|R_i = r) = p_{r(y_0,y_1)}$, as above:

$$Pr(Y_i(0) = 0|R_i = r) = [\underline{p}_{r(0,0)}, \overline{p}_{r(0,0)}] + [\underline{p}_{r(0,1)}, \overline{p}_{r(0,1)}]$$
$$Pr(Y_i(0) = 1|R_i = r) = [\underline{p}_{r(1,0)}, \overline{p}_{r(1,0)}] + [\underline{p}_{r(1,1)}, \overline{p}_{r(1,1)}]$$
$$Pr(Y_i(1) = 0|R_i = r) = [\underline{p}_{r(0,0)}, \overline{p}_{r(0,0)}] + [\underline{p}_{r(1,0)}, \overline{p}_{r(1,0)}]$$
$$Pr(Y_i(1) = 1|R_i = r) = [\underline{p}_{r(0,1)}, \overline{p}_{r(0,1)}] + [\underline{p}_{r(1,1)}, \overline{p}_{r(1,1)}]$$

That is, the bounds on the joint probabilities must be consistent with the marginal distributions of $Pr(Y_i(0) = y_0|R_i = r)$ or $Pr(Y_i(1) = y_1|R_i = r)$. If these constraints are binding such that the derived share bounds can be sharpened, this also further restricts the set of admissible covariances (since expectations are probabilities in the binary case), which may lead to further tightening of the estimated bounds. In practice, the relevance of this additional information is an empirical question.

Remark 2 (Extension to multiple treated potential outcomes): While we have outlined the mechanics to simultaneously condition on the treated and untreated potential outcome, it is straightforward to map the mechanics to condition on two *distinct* treated (or untreated) potential outcomes. For example, in the context of misdemeanor prosecution, conditioning on the likelihood of conviction *and* recidivism.

To see this, define the focal (treated) potential outcomes of interest as $(Y_i^j(1), Y_i^{j'}(1))$ for $j \neq j'$. We can then write $\pi_r(y_j, y_{j'})$ as

$$\pi_{r(y_j,y_{j'})} = E[D_i | R_i = r, Y_i^j(1) = y_j, Y_i^{j'}(1) = y_{j'}]$$

for $y_j \in \{0,1\}$ and $y_{j'} \in \{0,1\}$. Following the procedure outlined above will yield bounds on discrimination, conditional on treated potential outcomes $Y_i^j, Y_i^{j'}$. While the tight connection to conditioning on treatment effects no longer holds when conditioning on two treated potential outcomes, doing so nonetheless permits researchers to examine discrimination with richer conditioning sets and connect with more complex models of decision-maker behavior or notions of fairness.

B.2 Point identifying discrimination

This section discusses how to implement our discrimination estimation approach to obtain point estimates instead of bounds. We first sketch a simplified selection model commonly used to understand treatment non-compliance and treatment effect heterogeneity (Vytlacil, 2002). Individuals are treated if the benefit of treatment, $p_i(Y_i(1), Y_i(0), Z)$, outweighs the cost, where Z is a binary instrument that shifts the benefit of treatment. $p_i(Y_i(1), Y_i(0), Z)$ can also be interpreted as an individual's treatment propensity, and can be normalized such that $p_i(Y_i(1), Y_i(0), Z) \in [0,1]$. Let $u_i \in U[0,1]$ be a unidimensional measure summarizing possibly multiple factors that determine an individual's cost of treatment. Given the IV assumptions listed in Section 2.1, $Y_i(D_i)$ and u_i are unaffected by Z. Consolidating this notation, individual i is treated if $D_i = \mathbb{I}[p_i(Y_i(1), Y_i(0), Z) \geq u_i]$.

Returning to the discussion of always takers (A), compliers (C), and never takers (N) in Section 2.2, always takers have lower cost of treatment than compliers, who in turn have lower cost of treatment than never takers, $u_A \leq u_C \leq u_N$ (Angrist, Imbens, and Rubin, 1996).[46] Assuming that the relationship between treatment propensity and treated/untreated potential outcomes is linear, Equation 15 describes the expression for each marginal treatment response function, where $\bar{Y}_{(\cdot)}$ represents the average treated/untreated outcome for always takers, compliers, or never takers (Brinch, Mogstad, and Wiswall, 2017; Mogstad, Santos, and Torgovitsky, 2018; Kowalski, 2023b). The remaining steps to estimate discrimination follow Section 2.2, except that we estimate average treated or untreated potential outcomes by integrating the marginal treated or untreated functions ($MTO(p)$ or $MUO(p)$) over the full support of the treatment propensity. As a result, we obtain

[46] Vytlacil (2002) demonstrates how latent index selection models coincide with the local average treatment effect framework (Imbens and Angrist, 1994).

point estimates for the average treated/untreated outcomes and discrimination estimands.

$$
\begin{aligned}
MTO(p) &\equiv E[Y_i(1)|p = u_i] = \bar{Y}_{T,A} - \frac{p_A}{p_C}\left(\bar{Y}_{T,AC} - \bar{Y}_{T,A}\right) + \frac{2}{p_C}\left(\bar{Y}_{T,AC} - \bar{Y}_{T,A}\right) \times p \\
MUO(p) &\equiv E[Y_i(0)|p = u_i] = \frac{(2 - p_N)\bar{Y}_{U,NC} - (1 + p_A)\bar{Y}_{U,N}}{p_C} + \frac{2}{p_C}\left(\bar{Y}_{U,N} - \bar{Y}_{U,NC}\right) \times p \\
MTE(p) &\equiv E[Y_i(1) - Y_i(0)|p = u_i] = MTO(p) - MUO(p)
\end{aligned}
\tag{15}
$$

Empirical exercise: Racial discrimination in incarceration

We briefly demonstrate point identifying discrimination using the context of racial discrimination in incarceration decisions and publicly-available case-level records from Bexar County Criminal District (felony) Courts. We use a large reform meant to reduce overcrowding in Texas jails (SB 1067 in 1994). The goal of the reform was to reduce the burden on correctional facilities by limiting the incarceration rates for low-level offenders. The reform created a new category of felony: the state jail felony (SB 1067 Article 1, Subchapter C, §12.35) which reduced the punishment associated with a wide range of common offences, including many property and drug crimes. These provisions only applied to offences committed on or after September 1, 1994.[47]

Figure B1 validates this natural experiment separately for white and minority defendants. Panels a) and b) show that the reform resulted in a 6 p.p. (7%) increase in non-incarceration among white defendants and a 14 p.p. (20%) increase among minority defendants. Panels c) and d) demonstrate that future involvement with the criminal legal system falls by 3.1 p.p. for white defendants (–32.6%) and increases by 3.8 p.p. for minority defendants (29%), although the former estimate is imprecise. Panels e) and f) show that a summary measure of the baseline characteristics of defendants is smooth around this date cut-off.

Given that the reform is a valid natural experiment, we apply it to estimate the marginal treatment response functions as described above. We define treatment as $D_i = 1$ if an individual is not incarcerated (referred to as 'released' henceforth) and $D_i = 0$ if incarcerated. If an individual is released, we observe their treated re-offence outcome $Y_i(1)$. Here, we define $Y_i(1) = 1$ if an individual commits a new offence in Bexar County in the 12 months after they are released. Finally, while we assessed the validity of the natural experiment using regression discontinuity techniques, we parametrize Z as a binary instrument for simplicity: $Z = 1$ if an individual committed an offence after September 1, 1994.[48]

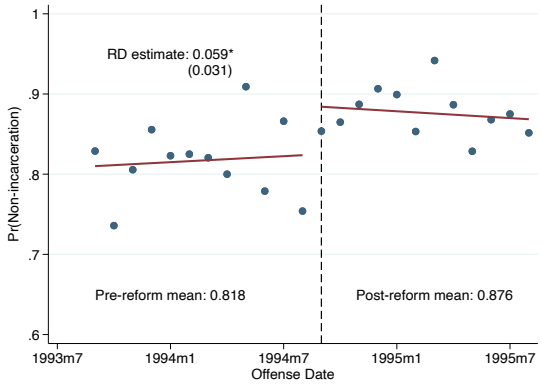
Figure B2 plots the estimated race-specific marginal treated outcome functions, $MTO_r(p)$ and Table B2 integrates these functions to estimate the average outcome that would be realized if everyone were released. The point estimates suggest large differences in underlying potential outcomes. 9.7% of white defendants would re-offend if released, while 17.7% of minority defendants would,

[47] Mueller-Smith and Schnepel (2021) use data from Harris County, TX to study a related aspect of the same legislation, which changed the incentives to offer deferred adjudication.

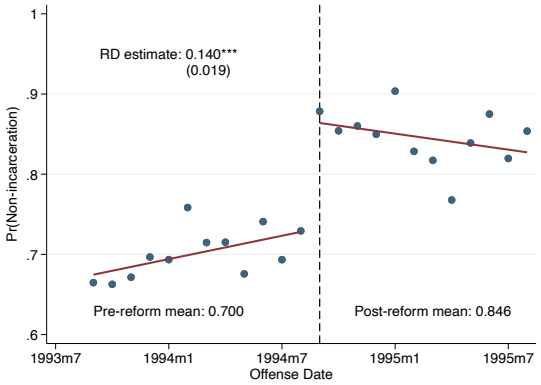
[48] We alternatively could have implemented this using RD variation as in Appendix B.3. Given the lack of trends in treatment and in re-offence outcomes here, we proceed with this binary implementation for simplicity.

Figure B1: Validating the legislative reform

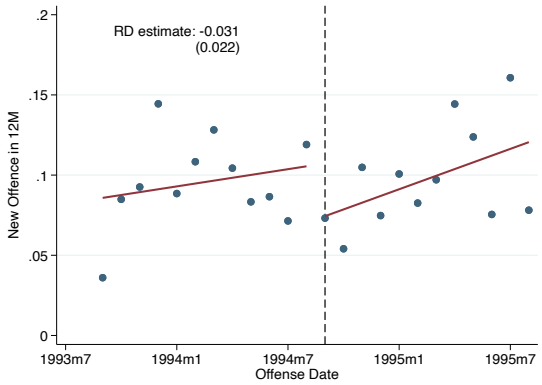
(a) Pr(Non-incarceration), white



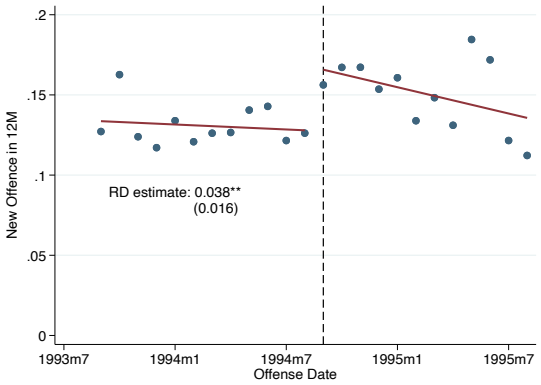
(b) Pr(Non-incarceration), minority



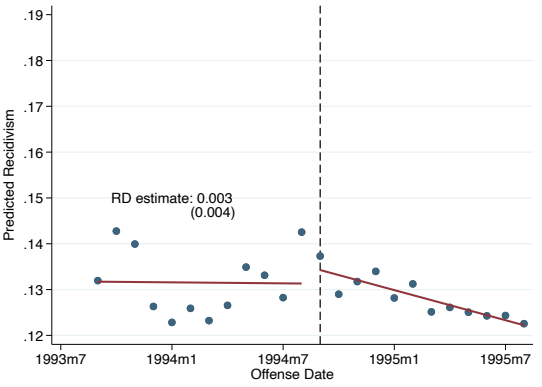
(c) Pr(New offence in 12M), white



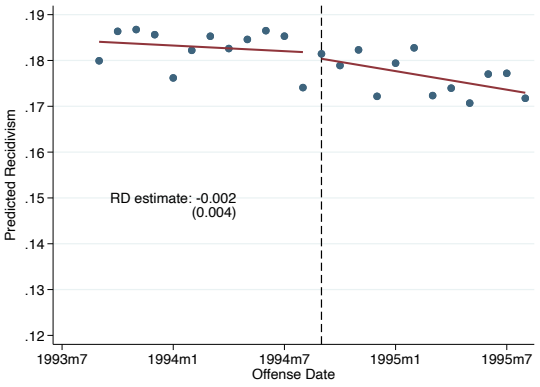
(d) Pr(New offence in 12M), minority



(e) Predicted 12M Recidivism, white



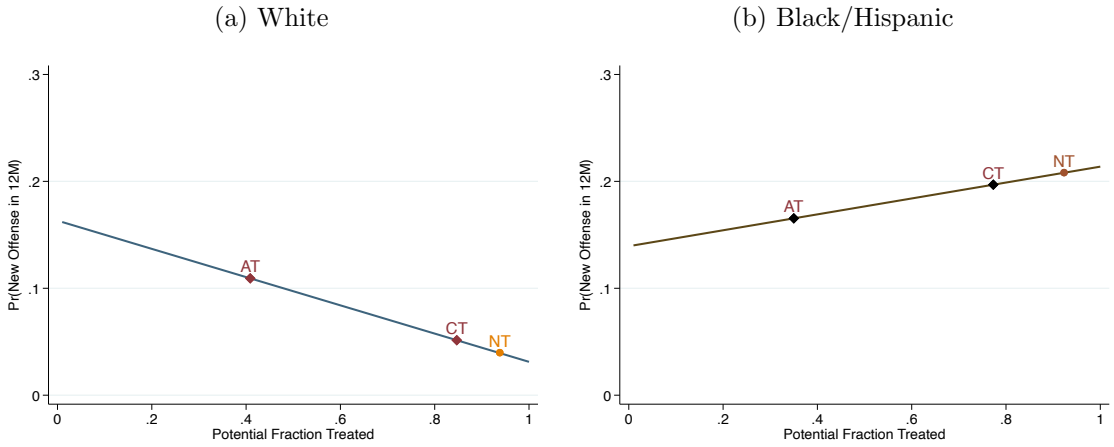
(f) Predicted 12M Recidivism, minority



Note: Each Panel presents RD estimates from regressions of the form $Y_i = \alpha + \beta \mathbf{1}(T_i > t) + \delta_1 T_i + \delta_2 \mathbf{1}(T_i > t) \times T_i + \varepsilon_i$, where T_i denotes the running variable, and t denotes the cut-off date of September 1, 1994. Sample includes all felony defendants who committed an offence in a year around the cut-off date. The lines of best fit are estimated on the monthly averages, represented by the blue dots. Incarceration is defined as serving an incarceration sentence. Predicted recidivism is computed by estimating $\mathbf{1}(\text{New Offence})_i = \alpha + \beta X_i + \nu_i$ using pre-reform data, and excluding the RD sample. These coefficients are then used to predict the probability of re-offending for the RD sample. X includes: indicators for race, offence type, felony category, gender, age, criminal history, and neighbourhood characteristics.

although these difference are not significant.

Figure B2: Average re-offence extrapolation by race



Note: This figure displays the marginal treated outcome curves ($MTO(p)$) (Mogstad, Santos, and Torgovitsky, 2018) for non-incarceration. The treatment is non-incarceration, or 'release', and the outcome is whether an individual commits a new offence within 1 year, if not incarcerated. Each $MTO(p)$ is identified by assuming a linear relationship between potential outcomes of always takers ('AT'), compliers ('CT'), and never takers ('NT') and their treatment propensities. Lower values of the x-axis denote individuals who are more likely to be released.

Table B2: Average re-offence estimates (p.p.)

	Average	White	Minority
	(1)	(2)	(3)
μ	0.155	0.097	0.177
95% CI	[0.121,0.188]	[0.011,0.180]	[0.143,0.211]

Note: This table presents estimates of the average treated outcome obtained using the approach described in Section 2.2 and Appendix B.2, separately by race. The treatment is non-incarceration and the treated outcome is whether an individual re-offends within one year after disposition. Confidence intervals are bootstrapped using 1,000 replications and a Bayesian bootstrap.

Following Equation 3, Table B3 displays point estimates of racial discrimination in non-incarceration decisions that condition on re-offence outcomes if not incarcerated. Prior to the reform, white individuals were 12.3pp more likely to be released than minority individuals. The reform significantly narrowed this disparity—after the reform, release rates were 3.4pp higher for White defendants, a 8.9pp reduction.

B.3 Measuring discrimination with a regression discontinuity (RD) design

We briefly illustrate how to use a regression discontinuity (RD) design to measure discrimination conditional on potential outcomes. We do this in the context of studying socio-economic discrimination in the decision to promote Michigan public school students to the next grade.[49] The

[49]This research used data structured and maintained by the MERI-Michigan Education Data Center (MEDC). MEDC data are modified for analysis purposes using rules governed by MEDC and are not identical to those data

Table B3: Estimated disparities (p.p.)

	Pre ($Z = 0$) (1)	Post ($Z = 1$) (2)	Change (3)
$\triangle$	0.123	0.034	-0.089
95% CI	[0.084,0.447]	[0.003,0.276]	[-0.170,-0.050]

Note: This figure presents the average disparities in each time period, conditional on treated potential outcomes, using the approach described in Section 2.2. The treatment is non-incarceration and the treated outcome, denoted by $Y_i(1)$, is whether an individual re-offends within one year after disposition. Confidence intervals are bootstrapped using 1,000 replications and a Bayesian bootstrap.

variation is generated by legislation that intended to improve the reading skills of 3rd graders in public schools. One component of the bill, also known as the 'Read by Grade 3' (RBG3) law, introduced a formulaic rule to determine when a student should be retained instead of being promoted. This component of the bill stipulated that 3rd graders scoring below 1253 (approximately the 5th percentile) on the standardized reading test (English Language Arts Michigan Student Test of Educational Progress, or ELA M-STEP) were to be retained while the rest were to be promoted to 4th grade. The promotion component of the policy came into effect in the 2020-21 school year but was ultimately repealed due to its unpopularity (Donahue, 2023; Povich, 2023). As a result, formula-based promotion and retention decisions affected students who were in 3rd grade during the 2020-21 and 2021-22 school years. As shown in prior work studying the impacts of the RBG3 policy, the formula induced shifts in promotion decisions and student outcomes at the cut-off (Westall et al., 2022a,b; Berne et al., 2023; Westall, Utter, and Strunk, 2023). We use these shifts to quantify whether there are differences by socio-economic status (SES) in the underlying potential outcome of promotion and adjust for them when measuring discrimination.

We use the same student-level data as used in prior work studying the RBG3 policy. Our analysis sample includes all first-time 3rd graders who scored within 15 points of the ELA M-STEP cut-off during the two years that the formula-based retention rule was active, which is the baseline sample in Berne et al. (2023). Students in our sample are more likely to be economically-disadvantaged, demonstrate limited English proficiency, and participate in special education programming than the average 3rd grader in Michigan's public schools.[50] We refer interested readers to the prior work for more details on these students' characteristics. For brevity, we refer to economically-disadvantaged students as 'low SES' and the rest of the students as 'high SES'.

Assessing the first stage & identifying assumptions

While this variation has been validated in aggregate by prior work, we assess the variation

collected and maintained by the Michigan Department of Education and/or Michigan's Center for Educational Performance and Information. This research was funded by a grant R305H1900004 through the U.S. Department of Education's Institute of Education Sciences, which is a collaboration between the University of Michigan and researchers from the Education Policy Innovation Collaborative (EPIC) at Michigan State University's College of Education. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not reflect the views of any other entity.

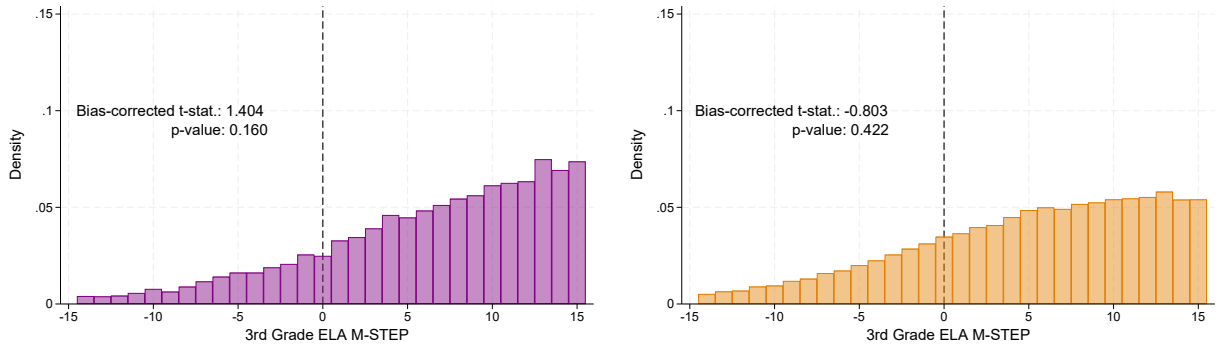
[50] A student is designated as economically-disadvantaged if the student: was eligible for free/reduced-price lunch, received SNAP/TANF, was homeless, was a migrant, or was in foster care.

separately by SES group since we implement our approach within these groups. Figure B3 finds no evidence of manipulation in the running variable, using the Cattaneo, Jansson, and Ma (2018) density test. Figure B4 presents a summary test of whether 4th grade ELA performance, predicted using baseline covariates, is smooth around the cut-off. While the covariates would predict that low SES students above the cut-off would score lower than low SES students below the cut-off, this estimate is marginally significant and economically small—the estimated discontinuity represents 0.005 of the standard deviation of the 4th grade ELA test score.

Figure B3: Smoothness of test score distribution density around cut-off

(a) High SES

(b) Low SES

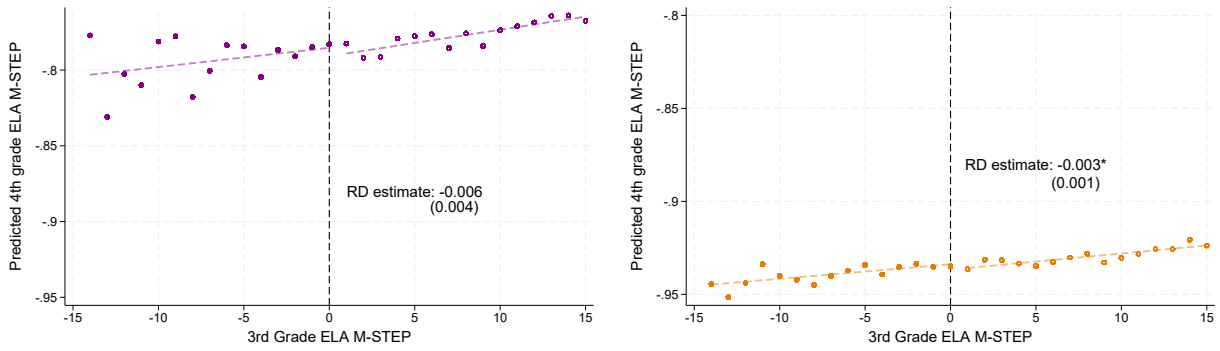


Note: These figures present results of the Cattaneo, Jansson, and Ma (2018) density test for the smoothness of the running variable. The x-axis represents the running variable, the 3rd grade ELA M-STEP, re-centered by the cut-off of 1252. The sample includes first-time 3rd graders during academic years 2020-21 and 2021-22 who scored within 15 points of the ELA M-STEP cut-off.

Figure B4: Smoothness of predicted 4th grade ELA M-STEP around test score cut-off

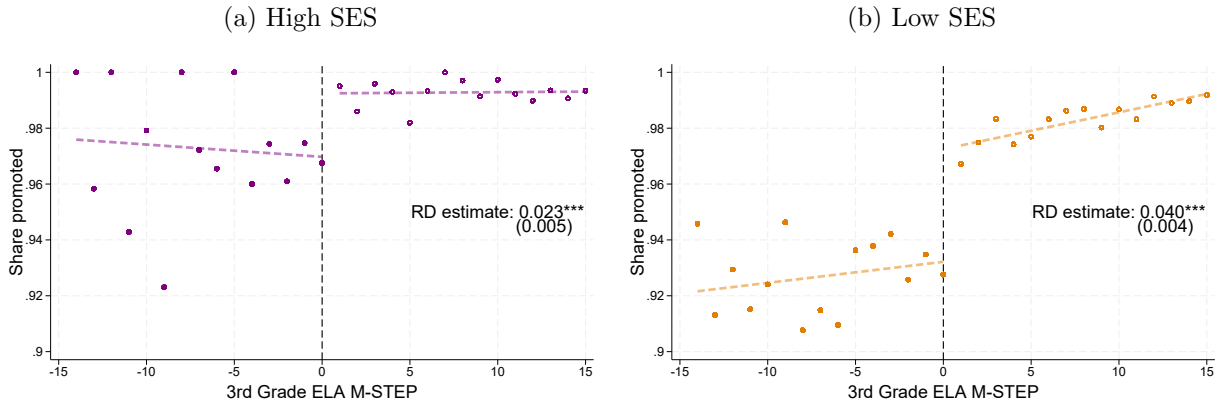
(a) High SES

(b) Low SES



Note: These figures present RD estimates investigating the impact of the RBG3 test score-based promotion policy on the predicted ELA M-STEP score taken in the following school year (regardless of actual promotion status), using demographics, Limited English Proficiency and special education status, whether the student was previously retained, whether the student is new to the district, and school fixed effects. 'RD estimate' presents β from $X_i = \alpha + \beta\,\mathbb{I}(\text{Score}_i{>}1252) + \delta_1 \text{Score}_i + \delta_2\,\mathbb{I}(\text{Score}_i{>}1252) \times \text{Score}_i + \varepsilon_i$. Standard errors are clustered at the level of the running variable. The x-axis represents the running variable, the 3rd grade ELA M-STEP, re-centered by the cut-off of 1252. The sample includes first-time 3rd graders during academic years 2020-21 and 2021-22 who scored within 15 points of the ELA M-STEP cut-off.

Figure B5: Effect of test score cut-off on promotion rates



Note: Each Panel presents RD estimates investigating the impact of the RBG3 test score-based promotion policy on promotion rates, using a local linear specification. The x-axis represents the running variable, the 3rd grade ELA M-STEP, re-centered by the cut-off of 1252. The sample includes first-time 3rd graders during academic years 2020-21 and 2021-22 who scored within 15 points of the ELA M-STEP cut-off. 'RD estimate' presents β from $\text{Promoted}_i = \alpha + \beta\,\mathbb{I}(\text{Score}_i{>}1252) + \delta_1\text{Score}_i + \delta_2\,\mathbb{I}(\text{Score}_i{>}1252) \times \text{Score}_i + \varepsilon_i$. Standard errors are clustered at the level of the running variable.

Figure B5 plots the relationship between the 3rd grade ELA M-STEP and the probability of being promoted, separately for high and low SES students. High SES students just above the cut-off are 2.3 p.p. (2.4% of the average promotion rate within 3 points of the cut-off) more likely to be promoted than high SES students below the cut-off. Low SES students just above the cut-off are 4.0 p.p. (4.3% of the average promotion rate within 3 points of the cut-off) more likely to be promoted than their counterparts below the cut-off. While the RBG3 law stipulated a formulaic approach to promotion and retention policy, these figures suggest that a large amount of discretion was still used in making promotion decisions, consistent with prior work studying the law's implementation (Westall et al., 2022a,b). Promotion rates do not jump from zero below the cut-off to one above the cut-off due to a section of the law that exempted certain students below the cut-off from being retained.[51] However, despite how common these exemptions seem to be, the test score cut-off still caused a modest but meaningful share of students who would have otherwise been retained to be promoted.

Estimating discrimination in grade promotion using the test score cut-off

Mapping the objects from the empirical discussion to the potential outcomes framework in Section 2.2, students are treated if they are promoted to 4th grade for the upcoming school year ($D_i = 1$) and untreated if retained in 3rd grade. Students are either high or low SES, denoted by $R_i = r \in \{h, l\}$. The treated potential outcome, $Y_i(1)$, is how well a student would perform in 4th grade standardized tests if promoted to the 4th grade for the upcoming school year. We construct our empirical analog of $Y_i(1)$ using students' test scores in the next school year, if they are promoted.[52] We do not observe how well students who are retained do in the 4th grade, since

[51] Common exemptions include students who: are English language learners, have disabilities, and whose parents submit an exemption request (Westall et al., 2022b).

[52] This definition of $Y_i(1)$ maps well to the underlying behavior of teachers and other educational staff. In fact,

they are still in the 3rd grade in the following year—as a result, we do not observe $Y_i(0)$. We define $Y_i(1) = 1$ if a student demonstrates "any proficiency" in both the Math and ELA M-STEP tests in the 4th grade, i.e., if they receive a score of at least Level 2 (out of 4) according to the Department of Education guidelines (Michigan Department of Education, 2023).[53] We consider this binary outcome to be a proxy for whether a promoted student was ready for the 4th grade. Hence $Y_i(1) = 1$ if a student is ready for 4th grade and $Y_i(1) = 0$ if a student is not ready. As discussed in Section 2.3, the instrument, Z, is an indicator for being above or below the RD cut-off.

Estimating shifts in the outcome if promoted

As before, we use quasi-experimental variation in promoted outcomes to estimate how average promoted outcomes vary across always takers, compliers, and never takers. We use that information to bound the average test score outcomes if everyone in the analysis sample were to be promoted. Figure B6 plots the relationship between the 3rd grade ELA M-STEP score and 4th grade outcomes. We limit the sample here to only students who were promoted, since we are trying to understand how promoted outcomes vary around the cut-off. Marginal high SES students who are promoted due to the RBG3 policy are 4.8 p.p. (35.9% of the average share proficient within 3 points of the cut-off) less likely to demonstrate 'any proficiency' in 4th grade than promoted students below the cut-off. Marginal low SES students are 2.2 p.p. (36.7% of the average share proficient within 3 points of the cut-off) less likely to demonstrate 'any proficiency' in the 4th grade than promoted students below the cut-off.

Following Section 2.3, we can use the intercepts of the local linear lines of best fit at the cut-off to estimate the proportions of always takers, compliers, and never takers as well as the average treated outcomes for always takers and compliers. As described in Section 2.2, this is the information we need to bound average treated outcomes for never takers and subsequently bound the average promoted outcomes if everyone **at the cut-off** from both SES groups were to be promoted.

Bounding the socio-economic group-specific average outcomes if promoted

Before presenting our SES-specific estimates of average outcomes if everyone were promoted, we first address the fact that the estimated discontinuities in outcomes if promoted are large relative to the first stage. For example, taken at face value, the estimates for high SES students suggest that a 2.3 p.p. increase in the probability of being promoted causes a 4.8 p.p. reduction in the probability of demonstrating any proficiency in 4th grade. The large magnitude of the discontinuity in outcomes could be due to noise or treatments other than promotion shifting around the cut-off.[54]

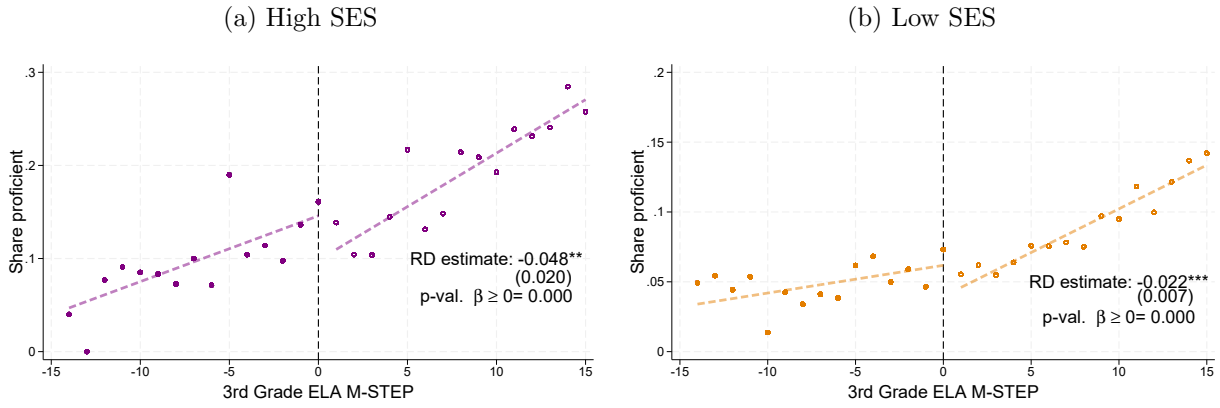
We deal with this issue by taking the sign of the discontinuity seriously but being agnostic about

educators were concerned that some students were advancing to 4th grade without the skills to cope and the reform was intended to improve skills that educators viewed as key inputs to students' success in later grades (French, 2019; Povich, 2023).

[53] We focus on a binary proficiency measure since our sample consists of students with 3rd grade ELA M-STEP scores around the 5th percentile of the score distribution. As a result, most of the variation in outcomes is between Levels 1 and 2.

[54] The policy mandated that students below the cut-off receive additional interventions and encouraged (but did not mandate) this for students above the cut-off. Berne et al. (2023) find evidence that some school districts provided the optional additional supports.

Figure B6: Impact of test score cut-off on 4th grade outcomes (Only promoted students)



Note: Each Panel presents RD estimates investigating the impact of the RBG3 test score-based promotion policy on 4th grade proficiency rates, using a local linear specification. 'Share proficient' represents the share of individuals who demonstrated any proficiency on both the Math and ELA M-STEP in 4th grade, as defined by Michigan Department of Education guidelines. The x-axis represents the running variable, the 3rd grade ELA M-STEP, re-centred by the cut-off of 1252. The sample includes first-time 3rd graders during academic years 2020-21 and 2021-22 who scored within 15 points of the ELA M-STEP cut-off, and were promoted to 4th grade. 'RD estimate' presents β from Any Proficiency$_i = \alpha + \beta\,\mathbb{I}(\text{Score}_i > 1252) + \delta_1 \text{Score}_i + \delta_2\,\mathbb{I}(\text{Score}_i > 1252) \times \text{Score}_i + \varepsilon_i$. The p-value in the second line is a one-sided test of whether the 'RD estimate' is weakly positive. Standard errors are clustered at the level of the running variable.

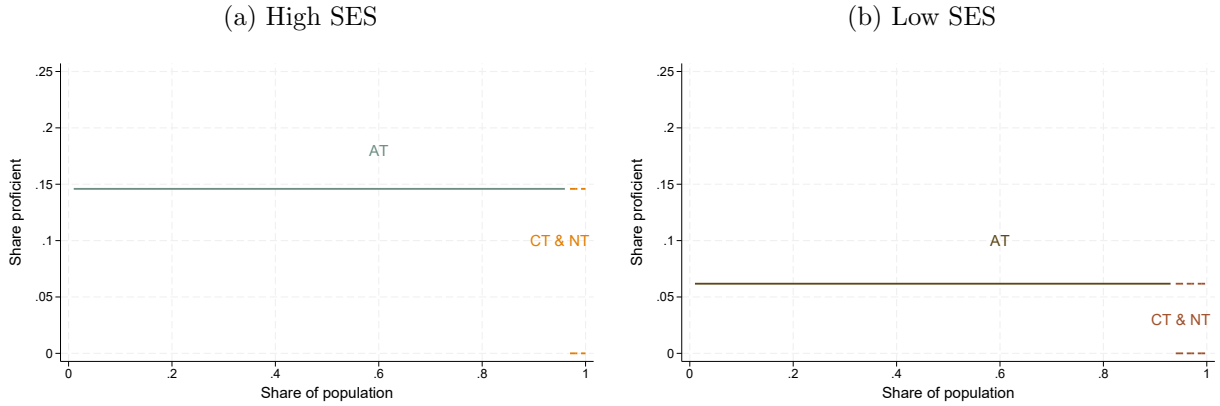
the magnitude. This approach assumes that marginally promoted students at the cut-off are weakly less likely to be ready for 4th grade than students who would have been promoted anyway. This is consistent with the literature finding that some 3rd grade students might benefit from retention in terms of academics (Jacob and Lefgren, 2004; Hwang and Koedel, 2022). The p-values presented in each Panel of Figure B6 suggest that our estimate of the sign of the relationship is credible—in both cases we reject the null that the relationship is weakly positive.[55]

We use this estimated **sign** of the discontinuity, rather than the **magnitude**, to adjust the bounding approach described in Section 2.2. Figure B6 implies that, for both SES groups, always takers for promotion are more likely to demonstrate 'any proficiency' in 4th grade than compliers. The approach described in Sections 2.2 and 2.3 would involve: i) backing out the proficiency rate if promoted for compliers and ii) assuming never takers' proficiency rate if promoted lies between zero and the compliers' proficiency rate if promoted. We modify this to assume that the average proficiency rate for both **compliers and never takers** together is bounded above by that of always takers, and below by zero. Under this adjustment, we still assume weak monotonicity of the relationship between compliance groups' treatment propensity and the average treated outcomes, using the estimated direction of the relationship between treated outcomes for always takers and compliers. Figure B7 displays the resulting estimates of average outcomes if promoted by compliance group and SES. Compliers and never takers, the portion of the population whose treated outcomes we bound, make up 3% and 6.8% of the population of high and low SES students

[55] We also conduct a placebo exercise that re-estimates the RD specification in Figure B6, using every other possible test score in our analysis sample as the cut-off, while keeping the bandwidth the same. Only 1.2% and 3.5% of placebo estimates for the high and low SES sample respectively are larger than the ones we observe, providing additional evidence validating the estimated discontinuity in treated outcomes.

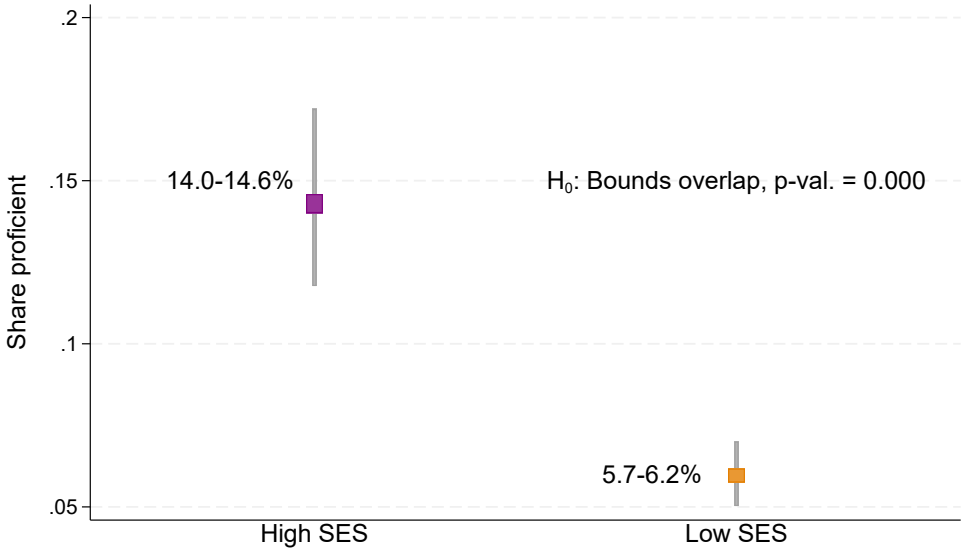
respectively.

Figure B7: Average outcomes if promoted ($Y_i(1)$) by compliance group



Note: This figure shows the average treated outcomes for always takers ('A'), compliers ('C'), and never takers ('N'). The treatment is promotion and the treated outcome, $Y_i(1)$, is whether a student demonstrated any proficiency on both the Math and ELA M-STEP in 4th grade, as defined by Michigan Department of Education guidelines. The bounds for the treated outcomes for never takers and compliers come from the assumption of weak monotonicity of average treated outcomes across compliance groups, and that $Y_i(1) \in \{0, 1\}$.

Figure B8: Average outcomes if promoted, $E[Y_i(1)|R_i = r]$



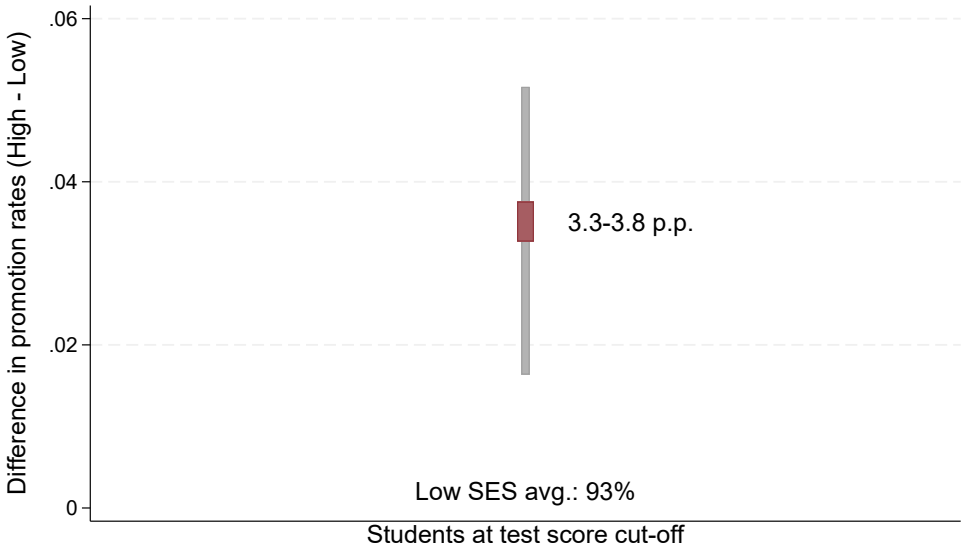
Note: This figure presents bounds on the average treated outcome obtained using the approach described in Section 2.2. The treatment is promotion and the treated outcome, $Y_i(1)$, is whether a student demonstrated any proficiency on both the Math and ELA M-STEP in 4th grade, as defined by Michigan Department of Education guidelines. Confidence intervals are for the true parameter and are bootstrapped using 1,000 replications and a Bayesian bootstrap. The p-value is from a formal bootstrapped test of whether the identified sets overlap, described in Appendix B.6.

With these adjustments, Figure B8 shows that there are indeed large SES differences in the underlying 4th grade outcomes that would be realized if all students in the analysis sample were promoted. 14.0–14.6% of high SES students would demonstrate 'any proficiency' if promoted to

4th grade, while this is true for only 5.7–6.2% of low SES students. Using a bootstrapped inference procedure described in Appendix B.6, we reject the null that these bounds overlap. Given that there are meaningful cross-SES differences in the underlying readiness for 4th grade, observed promotion gaps by SES are likely contaminated by differences in this unobservable factor.

Estimating the socio-economic promotion gap

Figure B9: SES promotion gap conditional on promoted outcome



Note: This figure presents bounds on the average difference in promotion rates conditional on treated potential outcomes, using the approach described in Section 2.2. The treatment is promotion and the treated outcome, denoted by $Y_i(1)$, is whether a student demonstrated any proficiency on both the Math and ELA M-STEP in 4th grade, as defined by Michigan Department of Education guidelines. Confidence intervals are for the true parameter and are bootstrapped using 1,000 replications and a Bayesian bootstrap.

Figure B9 presents our estimates of the SES gap in promotion rates for students who are at the test score cut-off. Our results indicate that high SES students are 3.3–3.8 p.p. (3.5–4.1%) more likely to be promoted relative to low SES students, even after accounting for SES differences in how prepared students are for the 4th grade.[56] Without our approach, an alternative strategy might be to measure promotion disparities that control for observable characteristics. 'Selection-on-observables' approaches that control for gender, special education status, English language learner status, race, school district characteristics, and neighborhood share with at least a BA would estimate high–low SES promotion gaps of 2.8 p.p., which is 15–26% smaller than the one that we find.

[56]The results are robust to weakening the assumption that despite any shifts in multiple treatments, the estimated sign of the discontinuity in promoted outcomes is still correct. We also estimate bounds for discrimination by allowing promoted outcomes for compliers and never takers to lie between 0 and 1. The resulting estimates imply that high SES students are 3.3–6.0 p.p. more likely to be promoted than low SES students. These patterns are in line with our baseline estimates, suggesting that estimates in Figure B9 are not driven by the assumption on the sign of the discontinuity in promoted outcomes.

Our analysis suggests that despite the intended formulaic nature of the RBG3 policy rules, the discretion that decision-makers exercised resulted in unwarranted disparities in promotion decisions by SES. These results make clear that the promotion disparities documented in recent work on the RBG3 law are not solely driven by differences in underlying unobservables (Westall et al., 2022a,b; Westall, Utter, and Strunk, 2023).

B.4 Extrapolating discrimination away from a regression discontinuity cut-off

Here we discuss the assumptions needed to apply the average potential outcome estimates derived from information at the cut-off to adjust treatment rates away from the cut-off. Let the analysis sample include the following values of the running variable (e.g., a test score): $s \in [\underline{s}, \bar{s}]$ and let s^* be the cut-off. Let D_i denote the treatment decision (e.g., whether student i is promoted) and Z denote whether a student is above or below the test-score cut-off. Consider that we want to condition on the treated potential outcome, $Y_i(1)$. Finally, any assumptions discussed below will need to hold by subgroup, although we omit subgroup notation for brevity.

Assume that the following assumptions hold:

RD1: $D_i \perp s$ and $E[D_i|Z=1, s^*] - E[D_i|Z=0, s^*] = E[D_i|Z=1, s] - E[D_i|Z=0, s] \; \forall \; s \in [\underline{s}, \bar{s}]$

RD2: $E[Y_i(1)|s = s^*] = E[Y_i(1)|s<s^*] = E[Y_i(1)|s>s^*]$

To understand the plausibility of these assumptions, consider them in the context of the student grade promotion application in Appendix B.3. RD1 states that within the analysis window, the test score does not influence promotion decisions by itself—only the cut-off does. RD1 also implies that the size of the first-stage would be the same in counterfactuals where the test score cut-off was elsewhere in the window. This ensures that the proportions of always takers, compliers, and never takers identified at the cut-off are applicable to the wider window. RD2 assumes that the average promoted outcomes that would be realized if all students **at the cut-off** were promoted is equal to the average promoted outcomes if students elsewhere in the window were all promoted. In the simplest case, this would be satisfied if $Y_i(1)$ did not vary across s.

While RD1 may hold in some settings, RD2 is a strong assumption that may not be satisfied in many contexts, including student grade promotion. To see this, note that we can split $E[Y_i(1)|s = s^*]$ into average treated outcomes for always takers, compliers, and never takers at the cut-off s^*. Let $p_{(\cdot)}$ denote the proportion of always takers (A), compliers (C), or never takers (N). Consider the first part of the equality in RD2, which compares students at the cut-off to those below. RD2 implies that the equality in Equation 16 must hold.

$$
\begin{aligned}
&E[Y_i(1)|s = s^*] = E[Y_i(1)|s<s^*] \\
&\implies p_A E[Y_i(1)|A, s = s^*] + p_C E[Y_i(1)|C, s = s^*] + p_N E[Y_i(1)|N, s = s^*] = \\
&\qquad p_A E[Y_i(1)|A, s<s^*] + p_C E[Y_i(1)|C, s<s^*] + p_N E[Y_i(1)|N, s<s^*] \\
&\implies p_C \left(E[Y_i(1)|C, s = s^*] - E[Y_i(1)|C, s<s^*]\right) + p_N \left(E[Y_i(1)|N, s = s^*] - E[Y_i(1)|N, s<s^*]\right) = \\
&\qquad = p_A \left(E[Y_i(1)|A, s<s^*] - E[Y_i(1)|A, s = s^*]\right)
\end{aligned}
\tag{16}
$$

We can map the implications of Equation 16 to how promoted outcomes vary with the running variable in the data. In Figure B6, all individuals below the cut-off are always takers for promotion (they are being promoted despite being below the cut-off). Here, always takers below the cut-off clearly have lower treated outcomes than those at the cut-off, implying that always takers **below the cut-off** are less likely to be prepared for 4th grade than always takers **at the cut-off**. Mathematically, this means that the right hand side of the final expression of Equation 16 is negative. For this assumption to hold with equality, the left hand side needs to be sufficiently negative to offset this. However, that would imply that compliers and never takers **at the cut-off** are **less prepared** for 4th grade than compliers and never takers **below the cut-off**, which is at odds with reasonable models of selection into promotion. Comparing the average treated outcomes at the cut-off to that above the cut-off yields a similar conclusion.

B.5 Identifying average potential outcomes with difference-in-difference designs

This section describes conditions under which we can estimate average potential outcomes in DiD settings with individual-level treatment non-compliance and heterogeneity in potential outcomes, focusing on non-parametrically bounding these quantities rather than obtaining point estimates. While 'treatment' in a typical DiD might be at an aggregate level, e.g., county-level, we consider a setting where individuals in either county may be treated both before or after some policy. The estimation of causal effects is confounded by the effects of time, which are correlated with the policy adoption. We use changes in the unaffected control county to purge these effects of time. Our approach is similar to "time-corrected" Wald approach to estimate the local average treatment effect (LATE) from recent work (De Chaisemartin and D'Haultfœuille, 2018). However, we require stronger assumptions to identify average potential outcomes for those who are not compliers.

We start with the following notation:

- $T \in \{0, 1\}$: Denotes periods before (pre) and after (post) a policy
- $G \in \{0, 1\}$: 1 if the county adopts the policy in $T = 1$, 0 if not.
- $Z \equiv T \times G \in \{0, 1\}$: This is the binary instrument

- $D_i(g,z) \in \{0,1\}$: Whether an **individual** takes up treatment or not.
- $Y_{it}(d,g)$: Potential outcomes for person i, given the time period, treatment status, and the county they are in.
- Refer to always takers, compliers and never takers as "compliance groups"

We make the following assumptions, many of which are common in IV implementations.

Assumption 1. First stage: $Pr(D_i(g,Z=1))>Pr(D_i(g,Z=0)) \ \forall \ g$

Assumption 2. Independence and exclusion: $(Y_{it}(1,g), Y_{it}(0,g), D_i(g,1), D_i(g,0)) \perp Z|g$. This implies that within each county, the instrument is random and only affects outcomes via changes in treatment status (Imbens and Angrist, 1994). This allows for a) potential outcomes to differ across counties and b) time-varying factors to directly affect potential outcomes.

Assumption 3. No spillovers: The potential outcomes of individual i are unrelated to the treatment status of other individuals (Angrist, Imbens, and Rubin, 1996).

Assumption 4. IV monotonicity: $D_i(g,1) \geq D_i(g,0) \ \forall \ g$

This allows the instrument to (weakly) shift individuals in only one direction across treatment contrasts and **does not allow secular trends to change treatment status of individuals**. Together, this means that only the following shifts between treatment contrasts are permitted:

1. $D_i(g,z) = 1 \ \forall \ z$: These are always takers in group g
2. $D_i(g,z) = 0 \ \forall \ z$: These are never takers in group g
3. $D_i(g,1) = 1$ and $D_i(g,0) = 0$: These are compliers in group g, shifted by the instrument

Finally, we make an additional assumption that is in the spirit of parallel trends assumptions, but is not typically made in IV settings or DiD estimation:[57]

Assumption 5. Parallel trends in potential outcomes: This assumes that i) the average change in treated outcomes is the same for always takers and compliers and is independent of county and ii) the average change in untreated outcomes is the same for never takers and compliers and is independent of county. This restricts the effects of time on potential outcomes to be constant across subsets of compliance groups, but not all of them, and does not force the effects of time to be identical across individuals.[58]

[57] This is similar to the assumption underlying the "time-corrected" Wald estimand in De Chaisemartin and D'Haultfœuille (2018). There, the treated (untreated) potential outcomes for those treated (not treated) in the pre-period are the same across group. Their assumption is enough to identify the LATE, but does not allow us to identify the average potential outcomes of each compliance group separately. It pins down time trends in a) treated outcomes for always takers and b) an average of untreated outcomes for both never takers and compliers. This does not pin down time trends for compliers specifically without further assumptions.

[58] This assumption is stronger when using our approach to obtain point estimates of average potential outcomes and discrimination rather than bounds. Obtaining point estimates requires assuming that the outcomes of always takers, compliers, and never takers are linearly related. Hence, if time trends in potential outcomes are assumed to be equal for two of the compliance groups, the assumption must extend to the remaining compliance group as well.

$$E[Y_{i1}(1,g) - Y_0(1,g)|g, \text{Always taker}] = E[Y_{i1}(1,g) - Y_0(1,g)|g, \text{Complier}] \text{ and } \perp g$$
$$E[Y_{i1}(0,g) - Y_0(0,g)|g, \text{Never taker}] = E[Y_{i1}(0,g) - Y_0(0,g)|g, \text{Complier}] \text{ and } \perp g$$

We now show how, under these assumptions, we can identify the proportions and average treated/untreated outcomes of always takers (A), never takers (N), and compliers (C) in the $G = 1$ county.

Given Assumption 4, there are no shifts in treatment status due to secular changes. Hence we have that the proportion of always takers (p_A), compliers (p_C) and never takers (p_N) in $G = 1$ are directly observed in the data for $G = 1$.

$$\begin{aligned} p_A &= E[D_i|G = 1, T = 0] \\ p_N &= 1 - (E[D_i|G = 1, T = 1]) \\ p_C &= 1 - (p_A + p_N) \end{aligned} \tag{17}$$

However, since common time trends can affect potential outcomes, the treated and untreated potential outcomes for each of these groups is not directly observed. To see this, recall that in settings where a binary instrument Z increases treatment take-up, the outcomes of individuals who are treated when $Z = 0$ identifies the treated outcomes for always takers. Equation 18 shows that if we try to estimate the treated outcomes for always takers in $G = 1$ using pre-period data (since $Z = 0$ & $G = 1 \implies T = 0$), we only recover treated outcomes for always takers in the pre-period. The difference between this and the treated outcomes for always takers in the post-period is the trend in treated potential outcomes, θ_1 (second line of Equation 18).

$$\begin{aligned} E[Y_i|D_i = 1, G = 1, Z = 0] &= E[Y_i|D_i = 1, G = 1, T = 0] = E[Y_{i0}(1,1)|A, G = 1] \\ \underbrace{E[Y_{i1}(1,1)|A, G = 1]}_{\text{Unobserved}} &= \underbrace{E[Y_{i0}(1,1)|A, G = 1]}_{\text{Observed}} + \underbrace{\theta_1}_{\text{Unobserved}} \end{aligned} \tag{18}$$

Equation 19 makes the same point for the untreated outcomes for never takers. In settings where a binary instrument Z increases treatment take-up, the outcomes of individuals who are not treated when $Z = 1$ identifies the untreated outcomes for never takers. Here, the outcomes of individuals who are not treated in the post-period only identifies the untreated outcomes for never takers in the post-period. Similarly, the difference between this and the untreated outcomes for never takers in the pre-period is the trend in untreated potential outcomes, θ_0 (second line of Equation 19).

$$E[Y_i|D_i=0,G=1,Z=1] = E[Y_i|D_i=0,G=1,T=1] = E[Y_{i1}(0,1)|N,G=1]$$
$$\underbrace{E[Y_{i0}(0,1)|N,G=1]}_{\text{Unobserved}} = \underbrace{E[Y_{i1}(0,1)|N,G=1]}_{\text{Observed}} - \underbrace{\theta_0}_{\text{Unobserved}} \tag{19}$$

Assumption 5 lets us identify the time trends in treated and untreated outcomes in $G=1$ (θ_1 & θ_0) using the change over time in $G=0$. Starting with treated outcomes, note that the only individuals who would be treated in the control county are always takers. The average change in treated outcomes in $G=0$ identifies the time trend for always takers in $G=1$, since Assumption 5 states that the trend in treated potential outcomes for always takers is identical across counties and is equal to the trend for compliers (see Equation 20).

$$\begin{aligned}\theta_1 = E[Y_{i1}(1,1) - Y_{i0}(1,1)|G=1,A] &= E[Y_{i1}(1,1) - Y_{i0}(1,1)|G=1,C] \\ &= E[Y_{i1}(1,0) - Y_{i0}(1,0)|G=0]\end{aligned} \tag{20}$$

Similarly, the individuals who are untreated in the control county, $G=0$, consist of compliers and never takers. From Assumption 5, the change in untreated outcomes for never takers and compliers are identical to each other, which means the average change in untreated outcomes in $G=0$ equals the average change in untreated outcomes for never takers in $G=0$. Additionally, Assumption 5 states that the time trend in untreated outcomes for never takers is identical across counties, which allows us to identify the time trend in untreated outcomes for never takers in $G=1$ (see Equation 21).

$$\begin{aligned}\theta_0 = E[Y_{i1}(0,1) - Y_{i0}(0,1)|G=1,N] &= E[Y_{i1}(0,1) - Y_{i0}(0,1)|G=1,C] \\ &= E[Y_{i1}(0,0) - Y_{i0}(0,0)|G=0]\end{aligned} \tag{21}$$

Equation 22 restates this by combining this with Equations 18 and 19, to show how the time trend in the treated outcomes of always takers and untreated potential outcomes of never takers in $G=1$ can be identified using the aggregate changes in treated and untreated potential outcomes in the control county, $G=0$.

$$\begin{aligned}E[Y_{i1}(1,1)|A,G=1] &= E[Y_{i0}(1,1)|A,G=1] + \theta_1 \\ &= E[Y_{i0}(1,1)|A,G=1] + E[Y_{i1}(1,0) - Y_0(1,0)|G=0] \\ E[Y_{i0}(0,1)|N,G=1] &= E[Y_{i1}(0,1)|N,G=1] - \theta_0 \\ &= E[Y_{i1}(0,1)|N,G=1] - E[Y_{i1}(0,0) - Y_{i0}(0,0)|G=0]\end{aligned} \tag{22}$$

We now have the treated outcomes for always takers and the untreated outcomes for never takers from $G = 1$ in both periods. We can use this information with other observed moments in the data to estimate the treated and untreated outcomes for compliers (Imbens and Rubin, 1997).

Starting with treated outcomes, the first line of Equation 23 notes that the observed outcomes among treated individuals in period $T = 1$ is a weighted average of treated outcomes for always takers and compliers in $T = 1$. Rearranging this expression, the second line shows that the treated outcomes for compliers in $T = 1$, $E[Y_{i1}(1,1)|C, G = 1]$, is a function of moments that we can estimate. $E[Y_{i1}(1,1)|A, G = 1]$ is obtained from Equation 22, $E[Y_i|D_i = 1, G = 1, T = 1]$ is a sample average, and each of the proportions, $p(\cdot)$, is obtained using Equation 17.

$$\begin{aligned} E[Y_i|D_i = 1, G = 1, T = 1] &= \frac{p_A E[Y_{i1}(1,1)|A, G = 1] + p_C E[Y_{i1}(1,1)|C, G = 1]}{p_A + p_C} \\ E[Y_{i1}(1,1)|C, G = 1] &= \frac{(p_A + p_C)E[Y_i|D_i = 1, G = 1, T = 1] - p_A E[Y_{i1}(1,1)|A, G = 1]}{p_C} \end{aligned} \tag{23}$$

We can estimate untreated outcomes for compliers in a similar way. The first line of Equation 24 shows that the observed outcomes among untreated individuals in period $T = 0$ is a weighted average of untreated outcomes for never takers and compliers in $T = 0$. Rearranging this expression, the second line shows that the untreated outcomes for compliers in $T = 0$, $E[Y_0(0,1)|C, G = 1]$, is a function of moments that we can estimate. $E[Y_0(1,1)|N, G = 1]$ is obtained from Equation 22, $E[Y|D_i = 0, G = 1, T = 0]$ is a sample average, and each of the proportions, $p(\cdot)$, is obtained using Equation 17.

$$\begin{aligned} E[Y_i|D_i = 0, G = 1, T = 0] &= \frac{p_N E[Y_{i0}(0,1)|N, G = 1] + p_C E[Y_{i0}(0,1)|C, G = 1]}{p_N + p_C} \\ E[Y_{i0}(0,1)|C, G = 1] &= \frac{(p_N + p_C)E[Y_i|D_i = 0, G = 1, T = 0] - p_N E[Y_{i0}(0,1)|N, G = 1]}{p_C} \end{aligned} \tag{24}$$

As a result, we have identified the following objects:

- Treated outcomes in each period for always takers
- Untreated outcomes in each period for never takers
- Treated outcomes in $T = 1$ for compliers
- Untreated outcomes in $T = 0$ for compliers

We are missing 2 objects: Treated outcomes in $T = 0$ for compliers and untreated outcomes in $T = 1$ for compliers. Assumption 5 allows us to recover this, since it implies the time trend in treated/untreated potential outcomes of each compliance group in $G = 1$ can be identified using

the aggregate changes in treated/untreated potential outcomes in the control county, $G = 0$. We now have average treated and untreated outcomes for each compliance group and in each period.

$$\begin{aligned} E[Y_{i0}(1,1)|C, G = 1] &= E[Y_{i1}(1,1)|C, G = 1] - \theta_1 \\ E[Y_{i1}(0,1)|C, G = 1] &= E[Y_{i0}(0,1)|C, G = 1] + \theta_0 \end{aligned} \tag{25}$$

B.6 Inference for tests of overlapping bounds

Here we discuss how we test whether the average potential outcome bounds estimated for each group overlap. Let the true parameter of interest for each group be μ_r, where $r \in \{m, w\}$ denotes the group. Let the estimated bounds be $[\mu_{r,L}, \mu_{r,U}]$. The goal is to test whether $[\mu_{m,L}, \mu_{m,U}]$ and $[\mu_{w,L}, \mu_{w,U}]$ overlap.

We construct a set that denotes the difference between the upper bound for one group and the lower bound for another: $\boldsymbol{M_d} \equiv [\mu_{m,L} - \mu_{w,U}, \mu_{m,U} - \mu_{w,L}] = [\widetilde{\mu}_L, \widetilde{\mu}_U]$. Note that $0 \in \boldsymbol{M_d}$ only if the bounds for each group are overlapping. To see this, consider the following three cases:[59]

Case 1. $[\mu_{m,L}, \mu_{m,U}]$ and $[\mu_{w,L}, \mu_{w,U}]$ are disjoint. E.g., $[\mu_{m,L}, \mu_{m,U}] = [0.5, 0.6]$ and $[\mu_{w,L}, \mu_{w,U}] = [0.2, 0.3]$. Then, $\boldsymbol{M_d} = [0.2, 0.4]$

Case 2. $[\mu_{m,L}, \mu_{m,U}]$ and $[\mu_{w,L}, \mu_{w,U}]$ overlap but one is not a subset of the other. E.g., $[\mu_{m,L}, \mu_{m,U}] = [0.25, 0.6]$ and $[\mu_{w,L}, \mu_{w,U}] = [0.2, 0.3]$. Then, $\boldsymbol{M_d} = [-0.05, 0.4]$

Case 3. $[\mu_{m,L}, \mu_{m,U}]$ is contained within $[\mu_{w,L}, \mu_{w,U}]$. E.g., $[\mu_{m,L}, \mu_{m,U}] = [0.2, 0.8]$ and $[\mu_{w,L}, \mu_{w,U}] = [0.3, 0.4]$. Then, $\boldsymbol{M_d} = [-0.2, 0.5]$

Our goal is to test the following null hypothesis: $H_0 : 0 \in \boldsymbol{M_d}$. We bootstrap the estimation of the bounds using a Bayesian bootstrap (Rubin, 1981). For each bootstrap replication, we construct the interval $\boldsymbol{M_d} \equiv [\mu_{m,L} - \mu_{w,U}, \mu_{m,U} - \mu_{w,L}] = [\widetilde{\mu}_L, \widetilde{\mu}_U]$. We then calculate a p-value as the share of the bootstrap replications in which $0 \in \boldsymbol{M_d}$, i.e., in which the bounds overlap.

[59] There are 3 more cases if you switch m and w, but they yield the same conclusions. These conditions also hold if the intervals themselves contain 0.