Measuring discrimination using natural experiments *

Nikhil Rao† James Reeves

November 4, 2024

Frequently updated: Click here for the latest version

Abstract

Disparities in high-stakes decisions are common, but difficult to interpret as discrimination if unobservable group differences exist. We show how to use natural experiments and a binary instrumental variable strategy to measure discrimination, adjusted for group differences in unobserved potential outcomes. Our approach does not require random assignment to decision-makers, a prerequisite for existing techniques. We study discrimination in two settings. First, we measure racial discrimination in misdemeanor prosecution with a budget cut that reduced prosecution rates in King County, Washington and a difference-in-difference strategy. Before the budget cut, we find no evidence of discrimination in prosecution conditional on unobserved potential recidivism. Afterwards, white defendants were more likely to be prosecuted than minority defendants. The gap is driven by prosecutors responding to the cut by dropping low quality cases, which were more common among minority defendants. These patterns suggest disparities were generated in prior stages of the criminal legal system, which prosecutors attenuated after the budget cut. Second, we study socio-economic discrimination in student grade promotion in Michigan public schools using a regression discontinuity design. Economically disadvantaged students near a test score cut-off were less likely to be promoted than non-disadvantaged students, even after accounting for differences in unobserved academic ability.

*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, Elisa Facchetti (discussant), Christopher Hollrah, Emily Horton, Brian Jacob, Amanda Kowalski, Steven Mello, Emir Murathanoğlu, Tyler Radler, Katherine Richard, Roman Rivera, Brock Rowberry, Damián Vergara, and Iris Vrioni for helpful suggestions. We are grateful to Kevin Cottingham at the Washington Administrative Office of the Courts for help accessing and working 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 for the student grade promotion analysis. 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 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. Rao gratefully acknowledges financial support from the Poverty Solutions Doctoral GSRA award and the Rackham One Term Dissertation Fellowship.

†University of Michigan: nikhrao@umich.edu (Job Market Paper)

1 Introduction

Disparities are common in many contexts where high-stakes decisions are made, e.g., in education, employment, health, and the criminal legal system. Measuring discrimination accurately, however, can be challenging due to unobservable factors relevant to the decisions (Becker, 1957; Aigner and Cain, 1977; Charles and Guryan, 2011).[1] The ideal measure of discrimination is one that compares individuals from two different groups who are otherwise similar, including in terms of relevant unobservable characteristics. In the canonical example of hiring discrimination, this involves measuring the hiring gap between people from two different groups who would be equally productive if hired. Since productivity is only observable upon hiring, measuring such a gap in practice is difficult.

Prior work measuring discrimination adjusted for unobservable factors uses random assignment to decision-makers to extrapolate unobserved outcomes (e.g., productivity of workers not hired) and directly condition on them (Arnold, Dobbie, and Hull, 2022).[2] There are many contexts where quantifying discrimination is policy-relevant and groups may be unobservably different, but individuals are not randomly assigned to decision-makers.[3] E.g., measuring socio-economic discrimination in education decisions is difficult since students are rarely randomly assigned to teachers and unobserved student ability may vary by socio-economic status (Paufler and Amrein-Beardsley, 2014).

In this paper, we show how to use natural experiments to measure discrimination in a treatment decision (e.g., being hired) when groups (e.g., race) are unobservably different and decision-makers are not randomly assigned. We map common forms of quasi-experimental variation, e.g., regression discontinuity or difference-in-difference, to a binary instrumental variable (IV) framework (Angrist, Imbens, and Rubin, 1996). To build intuition, consider the example of racial discrimination in hiring. First, we use the binary IV to estimate how average potential outcomes (e.g., productivity) vary across workers who would always be hired ('always takers') and marginal workers hired due to the IV ('compliers') within each racial group (Imbens and Rubin, 1997). Since we cannot observe the productivity of unhired workers of each racial group ('never takers'), we extrapolate their outcomes using behavioral assumptions from the marginal treatment effects literature (Brinch, Mogstad, and Wiswall, 2017; Mogstad, Santos, and Torgovitsky, 2018; Kowalski, 2023a). Such extrapolations recover the productivity of unhired workers, which lets us condition on unobserved productivity and estimate average discrimination among equally productive workers. In general, this approach uses a binary IV to estimate discrimination conditional on potential outcomes.

We implement our approach to study discrimination in two settings where individuals are not randomly assigned to decision-makers, using different forms of natural experiments. In our primary application, we study racial discrimination in misdemeanor prosecution in King County, Washing-

[1] We are agnostic on the source of such differences. These might be due to multiple factors, including preferences or discrimination prior to the decision of interest, i.e. 'systemic discrimination' (Bohren, Hull, and Imas, 2022).

[2] Outcome tests often also use random assignment to decision-makers (Arnold, Dobbie, and Yang, 2018). Such tests can detect bias against individuals at the margin of treatment (Becker, 1957; Anwar and Fang, 2006; Canay, Mogstad, and Mountjoy, 2024) but do not capture statistical discrimination (Hull, 2021) and do not quantify the magnitude of discrimination conditional on unobservable differences.

[3] When group differences are minimal, measuring discrimination does not require these adjustments (Goncalves and Mello, 2021; Harrington and Shaffer, 2023; Tuttle, 2023).

ton. Here, we use a difference-in-difference (DiD) design generated by a cut to the prosecutors' budget, which reduced the probability of being prosecuted—the treatment of interest. Unlike a regression discontinuity (RD) design, which has a familiar mapping to an IV framework, it is more challenging to map DiD and IV because time trends in potential outcomes can act as confounders (De Chaisemartin and D'Haultfœuille, 2018). We overcome this by developing an approach to de-trend potential outcomes, assuming i) treatment is stable over time and ii) a parallel trends assumption on potential outcomes.[4] This approach is applicable to traditional 2×2 DiDs as well as DiDs with staggered adoption. This allows us to map the DiD design to the binary IV method described above, estimate the distribution of potential re-offence outcomes by race, and subsequently estimate average discrimination adjusted for differences in potential outcomes.

In another application, we measure discrimination by socio-economic status (SES) in the decision to promote Michigan public school students to the next grade. We use an increase in promotion probability at a test-score cut-off and an RD design to estimate discrimination adjusted for SES differences in the unobserved success if promoted. While this analysis does not require the auxiliary DiD assumptions, the resulting discrimination estimates are valid only at the RD cut-off without additional assumptions (Angrist and Rokkanen, 2015; Cattaneo et al., 2021).

To outline the details of our approach, consider a potential outcomes framing of prosecution (Rubin, 1974). Define treatment as $D_i = 1$ if an individual is prosecuted, and $D_i = 0$ if dismissed. For simplicity, let the potential outcomes associated with each treatment state be binary.[5] The treated potential outcome, $Y_i(1)$, is whether an individual re-offends after being prosecuted. The untreated potential outcome, $Y_i(0)$, is whether an individual re-offends after being dismissed.

We start by defining discrimination as the racial gap in prosecution rates between individuals who would have the same re-offence outcome if prosecuted, $Y_i(1)$. We later discuss racial gaps between people with the same i) re-offence outcome if dismissed ($Y_i(0)$), or ii) treatment effect of prosecution ($Y_i(1) - Y_i(0)$), as well as empirical and conceptual reasons to prefer one over another. Since $Y_i(1)$ is binary in this example, the distribution of re-offence outcomes if prosecuted within each racial group is the share of individuals of each group who would re-offend if everyone were prosecuted. For each racial group, this is the "average prosecuted outcome".

The highlight the identification approach, assume that we can observe the average prosecuted outcome for each group. If the average prosecuted outcome varies by race, we quantify discrimination conditional on the prosecuted outcome in three steps. First, we rescale observed race-specific prosecution rates using the average prosecuted outcome—this yields race-specific prosecution rates that are conditional on prosecuted outcome. Second, we use the resulting prosecution rates to construct racial gaps in prosecution for each value of the prosecuted outcome. That is, we construct racial gaps in prosecution for people who: a) would re-offend after prosecution ($Y_i(1) = 1$) and b) would not re-offend after prosecution ($Y_i(1) = 0$). Third, the weighted average of these two racial gaps yields average discrimination conditional on the prosecuted outcome.

[4] This extends the logic of the "time-corrected" Wald estimator of local average treatment effects using DiD designs to estimate average potential outcomes instead (De Chaisemartin and D'Haultfœuille, 2018).

[5] We discuss conditioning on multi-valued potential outcomes in the sections below.

Since we cannot observe the average prosecuted outcome for each racial group, we estimate it using an IV framework. Consider a natural experiment that shifts treatment rates – prosecution rates in this example – and takes the form of a binary IV. Under standard IV assumptions, each racial group is partitioned into i) always takers, ii) compliers, and iii) never takers, and the average prosecuted outcomes of always takers and compliers can be estimated using the data (Imbens and Rubin, 1997). To identify average prosecuted outcomes for never takers, we restrict the relationship between likelihood of prosecution and average prosecuted outcomes. In our preferred approach, we assume this relationship is weakly monotonic across always takers, compliers, and never takers. For example, if compliers are more likely than always takers to re-offend if prosecuted, weak monotonicity implies that never takers must be at least as likely as compliers to re-offend if prosecuted (Kowalski, 2023a). This assumption yields bounds for the average outcomes that would be realized in a counterfactual where everyone in a given racial group were prosecuted, and hence yields bounds for average discrimination conditional on prosecuted outcome.[6,7]

Estimating discrimination with a DiD, as in our application where a budget reform affects only one county, requires additional assumptions. This is because it is difficult to disentangle the impacts of the reform from the effects of time, and to pin down the proportions and average outcomes of always takers, compliers, and never takers. We overcome this challenge by assuming that: i) time trends do not affect treatment (prosecution in this case) and ii) the average prosecuted outcome evolves similarly over time for always takers and compliers, and is independent of county.

The first assumption ensures that always takers, compliers, and never takers are defined as they would be with a binary IV, since only the reform (and not time) shifts treatment. Combined with the second assumption, the change in outcomes experienced by prosecuted individuals in the "no-reform" counties is a valid estimate of the change in prosecuted outcomes that always takers and compliers in the *county that adopted the reform* would have experienced if the reform had not occurred. In settings with staggered policy adoption, the second assumption needs to hold between each *county that adopted the reform* and all counties that do not or have not yet adopted the reform. This adjustment allows us to use DiD to estimate average prosecuted outcomes by race group, and estimate discrimination conditional on prosecuted outcomes as described above.[8]

In the context of misdemeanor prosecution, we estimate bounds for the race-specific average re-offence outcome if everyone were prosecuted using a large, unanticipated cut to the Prosecutor's Office budget in King County, Washington (Seattle and surrounding areas). Using administrative court records from Washington State, we employ a DiD strategy comparing King County to adjacent counties. The budget cut reduced prosecution rates by 20% and the likelihood of being charged

[6]Stronger assumptions on the relationship between selection into treatment and potential outcomes (e.g. assuming it is linear) point-identifies average potential outcomes. This approach involves estimating the set of marginal treatment response functions and integrating them (Mogstad, Santos, and Torgovitsky, 2018).

[7]Jordan (2024) imposes a mathematically similar restriction by modelling the objectives of felony review prosecutors who are randomly assigned cases. Our approach restricts the natural experiment instead. This allows decision-makers to adopt complex and multi-dimensional models as long as their decisions generate patterns that are consistent with the assumptions on the relationship between likelihood of treatment and average potential outcomes.

[8]We also discuss the assumptions required to use a DiD to estimate average untreated outcomes, and provide empirical validation for all of these assumptions when discussing the application.

with a new offence in the following year by 13–15%.[9] Reassuringly, we find no evidence that the budget cut affected other aspects of the King County criminal legal system or economy.

Using the shifts in prosecution rates and re-offence outcomes generated by the budget cut, we find meaningful racial differences in the unobserved prosecuted outcome. Prior to the budget cut, 24.5–29.3% of white defendants would commit a new offence if prosecuted, while the same is true for 32.3-37.1% of minority defendants.[10] These estimates imply that minority defendants in this context would be 3–12.6 p.p. (10.2–51.4%; $p = 0.004$) more likely to commit a new offence after prosecution than their white counterparts. Since there are racial differences in unobserved potential outcomes here, raw or covariate-adjusted racial gaps in prosecution, which do not account for these differences, would be biased.

We use the estimates of the average outcomes if everyone were prosecuted to estimate bounds for racial differences in prosecution, conditional on the prosecuted outcome. We cannot reject the null of no racial gap in prosecution rates before the budget reform. This changes after the reform—even though prosecution rates fall overall, white defendants were 1.3–4 p.p. (1.8–5.6%) more likely than minority defendants to be prosecuted, after conditioning on racial differences in the unobserved prosecuted outcome.[11] Alternative approaches to estimate discrimination that do not adjust for differences in potential re-offence outcomes are biased, e.g., the covariate-adjusted racial prosecution gap after the reform is 4.4 p.p., which is outside our estimated bounds.

A potential explanation for the relatively higher post-reform prosecution rate for white defendants is that minority cases might be backed up by weaker evidence, perhaps due to discrimination in pre-prosecution decisions, e.g., policing (Goncalves and Mello, 2021; Owens and Ba, 2021). Prosecutors may also be less likely to pursue weak, low quality cases when facing budget cuts, since resources are scarce, and prosecuting such cases likely requires greater resources. We should expect more pronounced racial gaps in a subset of cases that are likely weak, if prosecutors selectively drop weak cases due to the reform, and if weak cases are more common among minority defendants.

We test this explanation by using pre-reform data to classify offence based on the share of charges that are successfully sentenced, a proxy for case quality. We split our sample into 'high quality' (drug, DUI, property, and weapons violations) and 'low quality' offences (traffic, 'other', and violent offences), based on this proxy.[12] Repeating our discrimination estimation by subsample, we find that the post-reform racial gaps are driven by the 'low quality' subset of cases. Conditional on unobserved re-offence outcomes if prosecuted, white defendants after the reform are 1.6–4.6 p.p. more likely to be prosecuted than their minority counterparts, which is more muted in the 'high quality' subsample (0.2–1.1 p.p.).

[9] The direction and magnitude of these estimates are consistent with the impact of diversion and non-prosecution for low-level offences in other settings (Mueller-Smith and Schnepel, 2021; Agan, Doleac, and Harvey, 2023).

[10] We define this broad 'minority' out-group because Native Hawaiian and Pacific Islanders individuals also face disadvantage in Washington (Hu and Esthappan, 2017; Malott, 2024) and are a large proportion of non-white defendants. Our results are also robust to defining the out-group as Black & Hispanic individuals.

[11] Our results are qualitatively similar if we condition on the re-offence outcome if dismissed instead.

[12] Case quality may not be the only factor that varies between these two categories of offence types, e.g., prosecuting these offences might require different amounts of resources. However, such a split is still useful to understand types of cases that might be prioritized in the presence of fiscal constraints.

In prosecuting most cases before the reform, prosecutors were passing through any pre-existing disparities from prior stages of the criminal legal system (Harrington and Shaffer, 2024). However, the patterns above provide suggestive evidence that prosecutors shifted their focus to high quality cases after the budget cuts, and may have offset disparities generated in prior criminal legal stages. This behavior is consistent with recent work on how prosecutors can use discretion to attenuate discrimination from pre-prosecution decisions (Harrington and Shaffer, 2023).

Next, in the context of Michigan public schools, we estimate SES discrimination in the decision to promote 3rd graders to 4th grade (D_i). Here, we bound SES gaps in promotion rates for students who would achieve the same success if promoted to 4th grade ($Y_i(1)$), which educators say is a key concern underlying promotion decisions. Our measure of $Y_i(1)$ is whether students meet the state guidelines for being at least partially-proficient in English and Math standardized tests, if promoted to 4th grade. For brevity, we refer to such students as "succeeding" in 4th grade. We estimate whether underlying "success" in 4th grade varies by SES using an RD design generated by Michigan's "Read by Grade 3" law. Due to this law, the probability of being promoted discontinuously increased at a cut-off around the 5th percentile of the 3rd grade standardized test score distribution (Westall et al., 2022a,b; Berne et al., 2023; Westall, Utter, and Strunk, 2023).

We find large SES differences in the average 4th grade "success" rates that would be realized if all students at the test score cut-off were promoted. At the cut-off, which is at the low end of the test score distribution, only 6.1–6.5% of low SES students would "succeed" if promoted while 14.4–14.8% of high SES students would do so. These differences imply that low SES students in this context are 7.9–8.7 p.p. (55–59%; p =0.002) less likely to "succeed" if promoted to 4th grade than their high SES counterparts. Given these SES differences in underlying potential outcomes, raw or covariate-adjusted promotion gaps would provide biased estimates of discrimination.

We find evidence of discrimination in student grade promotion, even after accounting for SES differences in underlying "success" rates. Our bounds imply that high SES students at the cut-off are 3.4–3.7 p.p. (3.7–4%) more likely to be promoted than low SES students. Supplementary exercises i) show that this gap is due to the promotion of students who would not "succeed" if promoted, ii) show that it is not driven by differential parental involvement, and iii) discuss extrapolating estimates away from the cut-off to elsewhere in the analysis window. Crucially, alternative estimates of discrimination that do not adjust for differences in potential 4th grade success are biased—the covariate-adjusted disparity is 2.7 p.p., outside our estimated bounds.

This paper makes several methodological and empirical contributions. First, we add to the large body of literature on discrimination and its measurement (Becker, 1957; Aigner and Cain, 1977; Phelps, 1972; Arrow, 1973; Arnold, Dobbie, and Yang, 2018; Canay, Mogstad, and Mountjoy, 2024) by demonstrating how to use natural experiments that yield binary IVs to obtain bounds or point estimates of discrimination conditional on potential outcomes or treatment effects. Recent work estimating discrimination conditional on unobservable factors exploits continuous variation in treatment rates using random assignment to decision-makers (Arnold, Dobbie, and Hull, 2022). Using such an approach to examine discrimination within subsamples of the data, e.g., by time

period, can suffer from lack of power or a first stage within the subsamples. In contrast, such exercises will typically be more feasible with a binary IV approach. Additionally, random assignment to decision-makers in such settings is usually only conditional on certain covariates, e.g., courts, offence type, days of the week. Conditioning on covariates to ensure random assignment can introduce bias if the covariates themselves are generated due to some discriminatory behavior (Ayres, 2010). While our approach can accommodate such conditioning, it does not necessarily require it. As a final methodological point, we add to the literature mapping DiD to IV by showing how to use DiD variation to estimate average potential outcomes.

Second, we add to the understanding of criminal prosecution. Our results in Washington represent the first evidence of racial discrimination in misdemeanor prosecution that documents and directly accounts for unobservable racial differences. We find patterns consistent with recent work showing how prosecutors use their discretion to attenuate discrimination in earlier stages of the criminal legal system (Harrington and Shaffer, 2023; Jordan, 2024), rather than amplify them (Kutateladze and Andiloro, 2014; Rehavi and Starr, 2014; Tuttle, 2023). Consistent with the literature on the impacts of prosecution, our assessment of the King County budget cut suggests that non-prosecution for minor offences reduces future criminal activity.

Third, while there is descriptive evidence of disparities in student promotion decisions (Locke and Sparks, 2019; Moller et al., 2006), our analysis in Michigan public schools provides the first evidence of SES discrimination that adjusts for unobservable differences. Our results show that the promotion disparities documented in recent work on the "Read by Grade 3" law are not solely due to SES differences in unobservables (Westall et al., 2022b; Westall, Utter, and Strunk, 2023).

2 Discrimination estimands of interest

We begin with a general potential outcomes framework (Imbens and Angrist, 1994). Individuals are chosen for a binary treatment, D_i, as a function of the unobservable potential outcomes. Individuals realize the potential outcome associated with chosen treatment state, $Y_i(D_i)$. Each potential outcome is only observed if the associated treatment state is realized. That is, the treated potential outcome, $Y_i(D_i = 1)$, is only observed for individuals who are treated. Similarly, the untreated potential outcome, $Y_i(D_i = 0)$, is only observed for individuals who are not treated. 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.[13]

Our goal is to quantify discrimination defined as group differences in treatment rates for individuals with the same potential outcome. We begin by estimating differential treatment between individuals who would realize the same outcome **if they were treated**.[14] In Section 3, we also

[13]This framework is a general version of the framework described in Arnold, Dobbie, and Hull (2022). In their setting of bail reform, individuals vary in a latent unobservable Y^*, which is only observed among treated (released) individuals, while no outcome is observed for untreated (detained) individuals. Similar to Canay, Mogstad, and Mountjoy (2024), we consider a framework that applies to settings where untreated outcomes are also selectively observed among those not treated.

[14]This approach requires the researcher to specify the potential outcomes to condition on, $Y_i(D_i)$. It is possible

discuss conditioning on outcome if not treated and the treatment effect. Which is the most appropriate factor to condition on may depend on empirical, theoretical, and normative features of the particular context. The notion of discrimination conditional on potential outcomes, formally described in Definition 1 for the treated potential outcome, maps to classic notions of fairness and discrimination in economics that focus on whether people of two different groups are being treated differently, despite being identical in some objective but latent quality. For example, this notion would consider racial gaps in hiring rates between people who would be equally-productive if they were hired as discrimination (Aigner and Cain, 1977). This definition also encompasses multiple sources of discrimination (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).[15,16]

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

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

This definition of discrimination 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 $Y_i(1)$. One approach to estimate discrimination might be to compute the raw gap in treatment rates by group, $E[D_i|R_i = r_1] - E[D_i|R_i = r_2]$. However, this will differ from the quantity in 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 might be to compute the treatment gap, conditional on a set of observed covariates. This 'selection-on-observables' approach alters the interpretation of the discrimination test. Instead of measuring the extent of differential treatment between people who have the same potential outcome but different group membership, it moves towards a narrower test measuring how much two individuals with identical **observables** are treated differently **because** of their group identity, generating 'included variables bias' (Ayres, 2010). In general, even if such bias is small, controlling for covariates will typically not recover the measure of differential treatment, conditional on potential outcome, in Definition 1 unless the covariate used is perfectly correlated with $Y_i(1)$.

that the underlying decision-makers also value other factors that are not captured by the 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 our chosen $Y_i(D_i)$ maps to a well-defined notion of fairness.

[15]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 such differences as potential drivers of discrimination and investigate them in our empirical applications.

[16]Grossman, Nyarko, and Goel (2024) argue that comparisons as in Definition 1 are unsuitable for measuring disparate impact. Taking bail as a motivating example, they argue that a racial gap in release rates conditional on potential misconduct outcomes should not indicate discrimination since such a gap can arise when judges' release decisions are solely based on predictions of post-release misconduct. This argument implicitly prefers a definition of discrimination comparing individuals with similar predicted misconduct. While such a definition can be legally appealing, decisions based on predicted misconduct could generate disparate impact in practice, especially if prediction quality varies by race, and would be captured by the type of definition above.

In practice, groups often differ in terms of unobserved potential outcomes in many settings where measuring discrimination is of interest. For example, relevant to our primary application studying racial discrimination in misdemeanor prosecution, discrimination in prior decision points of the criminal legal system, e.g., policing, could generate cross-race differences in the underlying potential outcome distributions. In the context of our application studying socio-economic discrimination in students' grade promotion decisions, differential access to educational inputs by socio-economic status might generate group differences in skills (and thus potential outcomes). The estimand described in Definition 1 measures discrimination holding such upstream differences fixed. This measures the magnitude of unwarranted disparities arising in the decision process of interest.

We next demonstrate how to use a natural experiment to identify the estimand in Definition 1. Consider a quasi-random intervention that generates a binary instrument, Z. Assume Z satisfies the usual instrumental variables (IV) assumptions of relevance, independence, exclusion and monotonicity. In our main application, we use a difference-in-difference (DiD) approach to isolate quasi-random shifts in treatment. Time trends in potential outcomes and treatment status that are inherent in a DiD approach can act as confounders and make it difficult to use DiD variation in an IV set up. For ease of exposition, we abstract away from time trends for now and let Z be a binary instrument, where $Z \in \{0, 1\}$ denotes periods before and after some intervention. We first discuss identification and measurement of the discrimination estimated in Definition 1 using this binary instrument. We then address the adjustments required for the identification using a DiD approach in Section 3.

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

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

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

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 can also be used to measure changes in discrimination **due to the intervention**, $\triangle_{z=1} - \triangle_{z=0}$, the difference in discrimination before versus after some intervention.[17]

To understand how to estimate our main object of interest, note that the building blocks of

[17]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 (see Appendix C.1). Such cross-group comparisons can also suffer from the pitfalls of conducting marginal outcome tests with discrete instruments (Canay, Mogstad, and Mountjoy, 2024).

$\triangle_z$ are treatment rates that are conditional on period, group and treated outcome (π_{zry}). Since treated outcomes are not always 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 2 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 $Y_i(1) \perp Z$.

$$\begin{aligned} \pi_{zry} &\equiv E[D_i|Z=z, R_i=r, Y_i(1)=y] \\ &= \frac{E[Y_i(1)=y|Z=z, R_i=r, D_i=1] \times E[D_i|Z=z, R_i=r]}{E[Y_i(1)=y|Z=z, R_i=r]} \\ &= \frac{E[Y_i(1)=y|Z=z, R_i=r, D_i=1] \times E[D_i|Z=z, R_i=r]}{E[Y_i(1)=y|R_i=r]} \end{aligned} \tag{2}$$

Equation 2 highlights how to quantify the period-, group-, and potential-outcome specific treatment rates used to estimate discrimination conditional on having the same outcome if treated ($\triangle_z$). We need the following objects:

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

Each period-, group-, and potential-outcome specific treatment rate π_{zry} is a function of two moments directly observable in data, and one typically unobserved moment. Objects 1 and 2 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 represents the underlying share of individuals of group r who would experience a given value of the treated potential outcome. This object 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 are the same across race, we could interpret the raw observed group differences in treatment as discrimination. We also use $E[Y_i(1)=y|R_i=r]$, along with each group's shares of the population (p_r), to compute the underlying prevalence of treated potential outcome y in population: $E[Y_i(1)=y] = p_{r_1}E[Y_i(1)=y|R_i=r_1] + p_{r_2}E[Y_i(1)=y|R_i=r_2]$. We use this to aggregate each π_{zry} to estimate the discrimination conditional on having the same outcome if treated in a given period ($\triangle_z$) in Definition 2.

However, $E[Y_i(1)=y|R_i=r]$ is not typically directly observable in the data, unless there are unique institutional features such as random assignment to supremely lenient decision-makers who treat (almost) everyone (Arnold, Dobbie, and Hull, 2022; Baron et al., 2023; Reeves, 2023). There

are many settings where measuring discrimination is of interest, but individuals are not randomly assigned to decision-makers. For instance, criminal misdemeanor defendants in King County are not randomly assigned to prosecutors, and students in Michigan public schools are not randomly assigned to teachers. Measuring discrimination in prosecution or students' promotion decisions in these contexts is thus difficult if the distribution of potential outcomes varies by group.

Next, 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. If so, we adjust for such differences following Equation 2, and then measure discrimination that accounts for differences in potential outcomes.

3 Estimating discrimination using a natural experiment

In this section we discuss using a binary instrumental variable (IV) to understand if the distribution of potential outcomes varies by social group using insights from the IV and marginal treatment effects literature. First, we describe estimating the share of each group that would experience a given **treated** potential outcome if everyone of that group were treated. This is a key input to estimating discrimination that is conditional on treated potential outcomes. We then discuss accounting for complications that arise when using difference-in-difference (DiD) variation for this purpose. Finally, we discuss conditions required to account for group differences in **untreated** potential outcomes or **treatment effects** instead, which we also discuss in the context of applications that have meaningful treated and untreated potential outcomes.

3.1 Implementation with a binary instrument

We describe the framework in the context of racial discrimination in misdemeanor prosecution. Individuals belong to either 'White' or 'Minority' groups, denoted by $R_i \in \{w, m\}$, and the treatment decision is prosecution. 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 5. We start with a discussion of estimating racial differences in prosecution rates for individuals who would have the same outcome if prosecuted, i.e. treated. Later, we discuss conditioning on the outcome if dismissed, or the treatment effect of prosecution.

Individuals, indexed by i, are chosen for treatment, D_i. In the context of prosecution, this decision is made by a bundle of 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., $Y_i(D_i) \in \{0, 1\}$. We use binary potential outcomes for the rest of this section for simplicity. However, note that the expressions in Definition 2 and Equation 2 accommodate multi-valued potential outcomes.

Finally, let Z be a binary instrument that shifts the rate of prosecution. Let Z represent periods

before and after an unanticipated budget cut that sharply reduced prosecution rates. As described in the previous section, our application isolates the quasi-experimental variation using a difference-in-difference (DiD) strategy in which one county adopts a reform and the others do not. The time trends inherent in this approach make directly using an IV challenging. For ease of exposition, we first abstract away from time trends and discuss how a binary IV identifies the required moments. We then make adjustments for the DiD in the following subsection.

Recall that we need to estimate each period-, group- and potential outcome-specific treatment rate (π_{zry}) in Equation 2 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. Surmounting 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, $E[Y_i(1) = 1|R_i = r]$ is the average outcome that people in group r would realize if everyone in that group 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 IV assumptions listed in Section 2, the variation from the binary instrument Z partitions the population into three "compliance groups" (always takers, compliers and never takers) and identifies their associated proportions and certain average potential outcomes (Angrist, Imbens, and Rubin, 1996). Assuming that 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 our main application, the treatment is prosecution, and the reform decreases prosecution rates. Hence, p_A is the share of people who are prosecuted after the budget cut, and p_C is the change in prosecution rates due to the budget cut. Since always takers, compliers and never takers partition the population, 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 some of these groups. Since our initial focus is on treated potential outcomes, we concentrate on that potential outcome here. In our main application, the average treated outcome for always takers is the average outcome of people treated (i.e., prosecuted) after the budget cut. Note that the group of people treated before the budget cut consist only of compliers and always takers, since never takers are never treated. 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 recovers average treated outcomes for two of the three "compliance groups" that partition the population: always takers and compliers. However, we do not observe the average treated outcomes of never takers, since they are never treated. This is the final piece to estimate the average outcome that would be realized if everyone of each race group were treated, $E[Y_i(1) = 1|R_i = r]$.[18]

[18]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.

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. 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 might be the case if prosecutors were more likely to prosecute individuals who were unlikely to commit a new offence if prosecuted. 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", similar to 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, which pins down the lower bound for never takers' average outcomes.[19] Since $Y_i(1) \in \{0, 1\}$, the average treated outcomes for never takers is bounded above by one. Combining the bounds on the treated outcomes for never takers with the point estimates for the average treated outcomes for compliers and always takers yields bounds on the average treated outcomes in each group's subsample, $E[Y_i(1) = 1|R_i = r]$.[20]

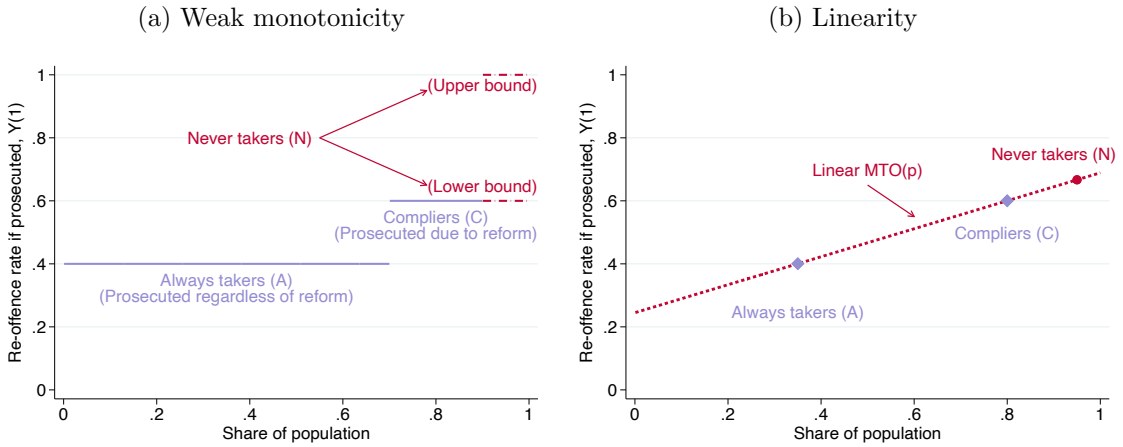
Note that the restrictions that we place to bound average treated outcomes are restrictions on the natural experiment and do not require assuming the underlying decision-makers focus on a single narrow objective. Rather, this approach allows decision-makers to adopt a wide range of complex and multi-dimensional models as long as their decisions generate patterns that imply that average potential outcomes for each of the compliance groups is weakly monotonic in their likelihood of being treated. In the context of prosecution, Panel a) of Figure 1 amounts to assuming that never takers, who are least likely to be prosecuted, are at least as likely as compliers to re-offend if prosecuted. This assumption might be violated if other inputs into prosecution decisions generate contradictory patterns. For example, assume prosecutors are also less likely to pursue cases with poor quality evidence, such that all the never takers' cases are poor quality. If individuals whose cases are poor quality do not tend to re-offend if prosecuted, this could violate our weak monotonicity assumption. In such situations, an alternative approach is to bound never takers' treated outcomes between 0 and 1—the widest logically possible bounds (Manski, 1989).

We obtain point estimates instead if we restrict the relationship between the underlying treatment propensity and treated outcomes to be linear. Panel b) of Figure 1 demonstrates this, where

[19] This restriction is mathematically similar to the "Performance Bound" in Jordan (2024), who studies racial discrimination in felony review in a context where prosecutors are randomly assigned cases. However, those bounds arise from modelling the underlying objectives of prosecutors, while ours place restrictions on the policy reform.

[20] If potential outcomes are multi-valued, we can still use this approach. 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 population prevalence and is hence also bounded between $[0, 1]$.

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.

the diamonds plot the treated outcomes for the median always taker and complier against their respective treatment propensities.[21] Assuming this relationship is linear allows us to extrapolate the treated outcomes across the support of the treatment propensity and point-identify the treated outcomes of never takers. This restriction also identifies the marginal treated outcome function ($MTO(p)$) which is integrated 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).[22] Since this approach involves assuming all of the marginal treatment response functions are linear, it places stronger restrictions than the previous partial identification approach.

Implementing this 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, $E[Y_i(1) = 1|R_i = r]$. This is the final object that we need to estimate discrimination that is conditional on treated potential outcomes, reproduced below for the case of binary treated outcomes in Equation 3.

[21] 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).

[22] Linearity assumptions identify all the marginal treatment response functions. We 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 $MTO(p)$, this identifies the marginal treatment effect function. Appendix C.2 sketches a simple model of selection, describes these assumptions, and illustrates this point-identification approach with a brief empirical example studying discrimination in incarceration decisions in Texas.

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

We observe the re-offence rate among prosecuted individuals in each period, Z 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 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).

There are important practical differences between our approach using natural experiments to estimate discrimination conditional on potential outcomes and prior work using random assignment to decision-makers (Arnold, Dobbie, and Hull, 2022). By using random assignment to decision-makers, prior work exploits continuous variation in treatment rates. As a result, exercises that examine discrimination within subsamples of the data, e.g., by time period, can suffer from lack of power or a first stage within the subsamples. In contrast, such exercises will typically be more feasible using our binary IV approach. Additionally, random assignment to decision-makers in such settings is usually only conditional on certain covariates, e.g., courts, offence type, days of the week. Conditioning on covariates to ensure random assignment can introduce 'included variables bias' if the covariates themselves are generated due to some discriminatory behaviour (Ayres, 2010). While our approach can accommodate such conditioning, it does not necessarily require it.

3.2 Accommodating difference-in-difference into the approach

So far, we have discussed using a binary IV to assess if treated outcomes differ by group (e.g., race) and estimate discrimination, conditional on treated outcomes. As discussed earlier, not all natural experiments easily map to the IV framework that is crucial for our approach. This is especially true for our main application with a DiD approach using the timing of a budget reform adopted

in one county but not others. Unlike typical DiD implementations, individuals are treated rather than counties and there is treatment non-compliance. That is, some individuals in counties that do not adopt a reform are still treated, and not all individuals in the county that adopts the reform are treated. Using a binary IV, e.g., before vs after the reform, in such a setting is complicated by time-variation, which does not let us disentangle changes in outcomes that are due to the reform from the effects of time.

We overcome this complication 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 in a counterfactual where it did not adopt the reform. This assumes that 1) time alone does not influence treatment status, and 2) that the time trend in average treated outcomes are the same for always takers and compliers, and are independent of county. These assumptions need to hold within the racial groups that we are trying to compare. Next, we introduce additional notation to formalize the assumptions and describe the adjustment. In keeping with the previous subsection, we follow the same example of measuring discrimination in prosecution, conditional on the prosecuted outcome using a budget cut that reduces prosecution rates. Appendix C.3 discusses the adjustment in further detail in a more general framework.

Let $T \in \{0,1\}$ denote periods before/after the reform, and let $G \in \{0,1\}$ denote the county that adopted the reform. Let $Z \equiv T \times G$ be an indicator for after the reform and in the county that adopted the reform. $D_i(g,z) \in \{0,1\}$ denotes whether an **individual** takes up treatment or not (i.e., is prosecuted or not). This is a key feature of the IV framework that differs from typical DiD implementations. Here, it is not the case that an entire county is "treated" by the budget reform. Rather, the budget reform shifts individuals into or out of treatment. As a result, individuals in either county can be treated both before or after the policy. The lack of a time subscript in the treatment indicator implicitly makes the first assumption—time does not influence treatment status. Similar to the monotonicity assumption in IV, we allow the reform to shift individuals into or out of treatment in only one direction.

This assumption allows us to partition the population of individuals in the county that adopts the reform ($G = 1$) into always takers (A), never takers (N) and compliers (C), since their proportions are constant over time. Equation 4 demonstrates how to estimate each of these in the data, following our main example where the reform reduces treatment rates.

$$\begin{aligned} p_A &= E[D_i|G=1,T=1] \\ p_N &= 1 - (E[D_i|G=1,T=0]) \\ p_C &= 1 - (p_A + p_N) \end{aligned} \tag{4}$$

Since time trends are still allowed to affect potential outcomes, the treated potential outcomes for each of these groups is not directly observed in both periods. In a setting where a reform reduces treatment take-up, the outcomes of individuals who are treated after the reform identifies the treated outcomes for always takers in $G = 1$. Equation 5 demonstrates that the treated

outcomes for always takers in the post-period and in the pre-period differ by the trend in treated potential outcomes, θ_1 (second line of Equation 5), which is unobserved.

$$\begin{aligned} E[Y_i|D_i=1,G=1,T=1] &= E[Y_{i1}(1,1)|A,G=1] \\ \underbrace{E[Y_{i1}(1,1)|A,G=1]}_{\text{Observed}} &= \underbrace{E[Y_{i0}(1,1)|A,G=1]}_{\text{Unobserved}} + \underbrace{\theta_1}_{\text{Unobserved}} \end{aligned} \tag{5}$$

We identify θ_1 by making an assumption in the spirit of parallel trends. Equation 6 formally describes this, where $Y_{it}(1,g)$ is an individual's treated potential outcome.[23] We assume that the average change in treated outcomes is the same for always takers and compliers, and is independent of county. This restricts the effects of time on treated potential outcomes to be constant across these two compliance groups, but not all of them, and does not force the effects of time to be identical across individuals.[24]

$$E[Y_{i1}(1,g)-Y_0(1,g)|g,\text{Always taker}] = E[Y_{i1}(1,g)-Y_0(1,g)|g,\text{Complier}] \text{ and } \perp g \tag{6}$$

Under these two additional assumptions, the change in outcomes among treated individuals in the counties that *did not adopt the reform* provides an estimate of the trend in treated outcomes in counties that *adopted the reform*: $\theta_1 = E[Y_{i1}(1,0)-Y_{i0}(1,0)|G=0]$. We estimate θ_1 and use it to separate the reform's impact on treated outcomes from the effects of time. This allows us to estimate average treated outcomes for always takers and compliers before and after the reform, in the county that adopted the reform. We implement these adjustments within race, and use the resulting moments to bound or point identify average outcomes for never takers before and after the reform, which lets us estimate the average outcomes if everyone of each racial group were treated.

The only difference between the DiD approach and the approach with a simple binary IV is that the average treated outcomes for each compliance group and the average outcomes if everyone were treated vary over time.As a result, bounds on average discrimination conditional on treated outcomes also vary with time.

The logic of this approach also extends to settings with staggered policy 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 assumption described

[23]This contrasts with the usual DiD implementations where individuals in counties where a policy occurs are considered 'treated' after the policy takes effect, while the rest are considered 'untreated' by the policy. There, the standard parallel trends assumption to identify impacts of the policy requires assuming parallel trends in the average untreated potential outcomes between the two counties.

[24]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. This identifies the LATE, but does not allow us to identify the average potential outcomes of each compliance group separately. That is because their assumption 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.

in Equation 6 needs to hold between each 'adopter' and its associated set of 'not-yet-adopters' in a window around policy adoption. Following the IV extrapolation approach described above and these DiD adjustments for each 'adopter' provides estimates of discrimination for each 'adopter'. This can then be aggregated across 'adopters' to construct an average measure of discrimination across all 'adopters', both before and after the policy.

3.3 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, or even the same **treatment effect**. Each of these may also map to different normative notions of fairness. For example, say we wanted to study discrimination in the decision to nominate students for advanced educational programs. Conditioning on the treated/untreated potential outcomes and the treatment effect in this example would provide an understanding of group differences in educational program nominations between individuals who would: i) do equally well in the program, ii) do equally well without the program, and iii) have equal gains from the program. We next discuss estimating these other discrimination estimands, and the additional assumptions that may be required, 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 outcomes is a function of period-, race- and potential outcome-specific treatment rates, as shown in Equation 7. 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.

$$E[D_i|Z=z,R_i=r,Y_i(0)=1] = \frac{E[Y_i(0)=1|Z=z,R_i=r,D_i=1]\times E[D_i|Z=z,R_i=r]}{E[Y_i(0)=1|R_i=r]} \quad (7)$$

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 7 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 computing 1 minus the **dismissal rate** for those who would have that outcome if dismissed. Appendix C.4 shows this formally. 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 binary IV, Appendix C.3 discusses the assumptions required to accommodate DiD variation.

Conditioning on treatment effects, $Y_i(1) - Y_i(0)$

Equation 8 describes treatment rates that condition directly on the treatment effect (τ_i). This object departs from Equation 3 in two ways. 1) The denominator is the share of individuals in the population who would realize a given treatment effect value. 2) The first term in the numerator is the share of treated individuals who would realize a given treatment effect value. We cannot estimate these quantities in the same way that we have done when conditioning on potential outcomes because τ_i is never observed for any individual and can take multiple values. For example, even if both $Y_i(1)$ and $Y_i(0)$ are binary, $\tau_i \in \{-1, 0, 1\}$. Thus even if we can estimate the local average treatment effect for compliers, we will not know the prevalence of a given value of τ_i.

$$E[D_i|Z = z, R_i = r, \tau_i = y] = \frac{E[\tau_i = y|Z = z, R_i = r, D_i = 1] \times E[D_i|Z = z, R_i = r]}{E[\tau_i = y|R_i = r]} \tag{8}$$
$$\tau_i \equiv Y_i(1) - Y_i(0)$$

Conditioning on the treatment effect requires restrictions on the support of the treatment effect. If we assume τ_i is binary, say due to some theoretical or empirical justification, we can overcome the two challenges above. For example, we may be in a context where the treatment (prosecution) is unlikely to reduce future re-offending. We then might assume $\tau_i \in \{0, 1\}$, which implies that the average treatment effect (ATE), $E[\tau_i|R_i = r]$, coincides with the denominator of Equation 8. This solves the first challenge. Similarly, $\tau_i \in \{0, 1\}$ implies that $E[\tau_i = y|Z = z, R_i = r, D_i = 1]$ is the average treatment effect on the treated (ATT). Recall that we directly observe average treated outcomes for the treated in the data and that the discussion regarding conditioning on untreated outcomes provided an estimate of the average untreated outcomes for the treated. The difference between these two averages is the ATT. Under the restriction that $\tau_i \in \{0, 1\}$, this identifies the proportion of treated individuals with a treatment effect of $\tau_i = 1$: $E[\tau_i = 1|Z = z, R_i = r, D_i = 1]$.

Plugging in bounds/point estimates from each of these steps into Equation 8 yields bounds/point estimates for $E[D_i|Z = z, R_i = r, Y_i(0) = y]$.[25] These treatment effect-specific treatment rates would then be aggregated up analogous to Equation 3, to estimate group differences in treatment among those who would have the same treatment effect, assuming binary treatment effects.

[25] Bounding this requires bounding a non-linear function of point identified and partially identified objects. Appendix C.4 discusses how we compute the bounds in such a case.

3.4 Recap: Assumptions and 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 group differences in the distribution of potential outcomes.[26] Implementing this requires the following:

1. A natural experiment that can be mapped to an IV framework
    - If using DiD variation, need time to not shift treatment, and a parallel trends assumption for potential outcomes
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 distribution of potential outcomes differ between the groups. For example, if $Y_i(D_i)$ is binary, this amounts to comparing the average outcomes if everyone of each group were treated (or not treated).
4. If the potential outcomes differ by group, estimate average discrimination conditional on potential outcomes following Equation 2.

We apply this approach to two empirical applications. The first studies socio-economic discrimination in student grade promotion, using a regression discontinuity approach. The second studies racial discrimination in prosecution, using a DiD approach. In both applications, 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 by group, and then plug these bounds directly into into the expression for our object of interest: the average period-specific discrimination, $\triangle_z$, in Equation 3. We can also estimate bounds on the underlying group- and potential outcome-specific treatment rates, which we report in appendices.[27]

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 the moments in the estimation procedure, enforcing the weak monotonicity restriction within each reweighted bootstrap sample. We then report the confidence intervals for the single true underlying parameter using the resulting bootstrap distribution (Imbens and Manski, 2004).

[26] 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 $j-$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.

[27] 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.

4 Socio-economic discrimination in student grade promotion

In this section we use the approach described in Section 3 to measure socio-economic (SES) discrimination in the decision whether to promote Michigan public school students to the next grade. This is a key decision in students' lives that can have large academic and non-academic impacts (Jacob and Lefgren, 2004, 2009; Eren, Lovenheim, and Mocan, 2022). We test for SES differences in potential outcomes if promoted using a formulaic rule that determined promotion decisions as a function of standardized test scores and generated a regression discontinuity (RD) design. Using the RD, we find large SES differences in the underlying chances of potential success if promoted to the next grade. We find significant SES gaps in promotion rates, even after accounting for SES differences in potential success in the next grade. We also highlight the relative advantages and disadvantages of implementing our approach with an RD design, which has a close link to the instrumental variables framework, rather than a difference-in-difference design.

4.1 Natural experiment: Michigan's 'Read by Grade 3' Law

In 2016, the Michigan legislature passed legislation regarding the retention and promotion of 3rd graders in public schools (Public Act 306 of 2016). The new bill, also known as the 'Read by Grade 3' (RBG3) law, intended to improve public school students' reading skills. One component of the law introduced a formulaic rule to determine when a student should be retained instead of being promoted. Prior to the bill, promotion decisions were at the discretion of the school or district staff. Now, 3rd graders scoring 1252 or lower (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.[28] The promotion component of the policy was meant to come into effect in the 2019-20 school year but was delayed by a year due to the COVID-19 pandemic, and eventually repealed since it was widely unpopular (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.

The formula-based system to decide promotion and retention for students presents an opportunity to study SES discrimination in student grade promotion decisions if the formula induced an exogenous shift in the probability of being promoted – i.e., the treatment rate – at the specified 3rd grade ELA-M-STEP cut-off score. We need this shift to quantify whether there are differences by SES in the underlying potential outcomes of promotion. We follow the approach in Section 3, but use a regression discontinuity (RD) design to estimate discrimination between students who would have the same potential outcome. Other quasi-experimental strategies that rely on the variation in the timing of these policies, e.g., difference-in-difference, would be challenging, given how close this policy was to the start of the COVID-19 pandemic.

We use the same data as in recent work studying the impacts of multiple aspects of the RBG3

[28] The policy also required schools to provide students below the cut-off with additional interventions. Appendix A.1 discusses this further, along with other details on the policy.

policy with an RD design at this test score cut-off. (Westall et al., 2022a,b; Berne et al., 2023; Westall, Strunk, and Utter, 2023; Westall, Utter, and Strunk, 2023). For each public school 3rd grader in the affected cohorts, we observe: their 3rd grade test scores, whether they were promoted to 4th grade, their performance in the following academic year (regardless of the promotion decision), and a host of baseline characteristics. We include all first-time 3rd grade ELA M-STEP test-takers during the two years that the formula-based retention rule was active. In our analysis sample, we only include students who scored within 10 points of the ELA M-STEP cut-off.[29]

Column 1 of Table 1 compares the general 3rd grade student body in the two affected cohorts to our analysis sample (Column 2). Michigan public school 3rd graders are diverse in terms of race, ethnicity and economic status. Overall, 32% of students are non-white and 54% of students are designated as facing some economic disadvantage.[30] We refer to economically-disadvantaged students as 'low SES' and the rest of the students as 'high SES'.

Since our analysis sample limits to students with ELA M-STEP scores in a small window around the 5th percentile cut-off, students here are relatively more disadvantaged than the average 3rd grader in Michigan's public schools. 82% are economically-disadvantaged, and 54% are non-white. Students in our analysis sample are more likely to attend a charter school, demonstrate limited English proficiency and participate in special education programming. Their baseline 3rd grade ELA M-STEP scores are lower by construction, 1255 on average (compared to 1293 among all 3rd graders). This places the average student in our sample at the lowest 3rd grade proficiency level designated by Michigan Department of Education ('Level 1: Not proficient').

Table 1: Characteristics of Michigan 3rd grade students (2020–22)

	Overall	Estimation sample
	(1)	(2)
N	163,412	21,790
Demographics		
White (Non-Hispanic)	0.68	0.46
Black	0.19	0.40
Hispanic	0.08	0.10
Male	0.51	0.56
Economic disadvantage	0.54	0.82
Academic		
3rd Grade M-STEP ELA score	1292.8	1254.8
Attends charter school	0.12	0.20
Limited English Proficiency (LEP)	0.08	0.12
Special Education	0.14	0.28

Note: Column 1 includes all Michigan public school 3rd graders who took the ELA M-STEP for the first time in academic years 2020-21 and 2021-22. Column 2 limits to those 3rd graders who scored between [1242, 1262] on the 3rd grade ELA M-STEP.

[29] We use a smaller bandwidth than other work studying RBG3 because we implement exercises that use objects estimated at the cut-off to inform discrimination away from the cut-off (but within the bandwidth). These require assumptions that are more likely to hold in a small window.

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

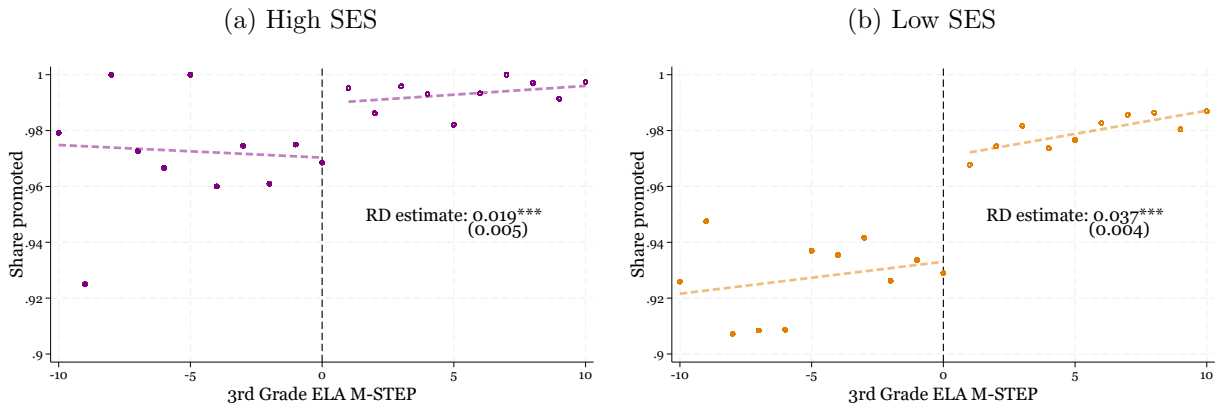
Despite the fact that this variation has been studied and validated by prior work, we present additional validation results since we use an analysis bandwidth that is smaller than prior work. We conduct all validation tests within SES, since we need a valid RD design for each SES. Figure A1 finds no evidence of manipulation in the running variable, using the Cattaneo, Jansson, and Ma (2018) density test. Figures A2–A4 show that demographics, academic characteristics, and predicted 4th grade performance are smooth around the cut-off.

Next, we turn our attention to whether the policy rule actually resulted in discontinuous changes in the probability of promotion. If so, we can use this to estimate average potential outcomes for always takers, compliers, and never takers, and use that information to estimate average outcomes if all students of both SES groups were promoted. We start by estimating Equation 9:

$$\text{Outcome}_i = \alpha + \beta\,\mathbb{I}(\text{Score}_i{>}1252) + \delta_1 \text{Score}_i + \delta_2\,\mathbb{I}(\text{Score}_i{>}1252) \times \text{Score}_i + \varepsilon_i \tag{9}$$

Figure 2 plots the relationship between the 3rd grade ELA M-STEP and the probability of being promoted, along with the estimated β from Equation 9, separately for high and low SES students. We see significant changes in the probability of being promoted around the cut-off for both groups of students. High SES students just above the cut-off are 1.9pp ($\approx 2\%$) more likely to be promoted than high SES students below the cut-off. Low SES students just above the cut-off are 3.7pp ($\approx 4\%$) more likely to be promoted then their counterparts below the cut-off.

Figure 2: 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-centred by the cut-off of 1252. The sample includes students who took the 3rd grade ELA M-STEP for the first time between 2020 and 2022 and scored within 10 points of the cut-off. 'RD estimate' presents β from Equation 9. Standard errors are clustered at the level of the running variable.

Notably, 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). The reason we do not see promotion shares jump from zero below the cut-off to one above the cut-off is due to a section of the law that allowed students below the cut-off (who should have been retained

according to the law) to be promoted if they met specific criteria.[31] 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.

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

We start by mapping the objects from the empirical discussion to the potential outcomes framework in Section 3. Students are considered treated if they are promoted to 4th grade for the upcoming school year ($D_i = 1$) and 'untreated' if retained in 3rd grade. The treated potential outcome, $Y_i(1)$, is how well a student would perform if promoted to the next grade for the upcoming school year. For students who are retained, we do not observe how well they do in the 4th grade, since they are still in the 3rd grade.[32] As a result, we do not observe $Y_i(0)$.

The instrument, Z, is an indicator for being above or below the RD cut-off (we discuss complications due to the local nature of the RD below) and we use the RD to assess if there are SES differences in how ready students are for the 4th grade. We then adjust for such SES differences following Section 3, and estimate SES promotion gaps for students who would have performed equally-well if promoted. We define whether students are high or low SES by whether they are flagged as being economically-disadvantaged, and denote this by $R_i = r \in \{h, l\}$.

We construct our empirical analog of $Y_i(1)$ using students' test scores in the next school year, if they are promoted. Since Michigan public school students in 4th grade are required to undergo standardized testing, we observe an objective measure of student performance among the promoted students when they take the 4th grade test. This has a direct mapping to fairness norms—between two students equally likely to do well in 4th grade, it would be unfair to promote one student and not the other. This also maps well to the underlying decision problem followed by teachers and other education staff. In fact, there were concerns that some students were advancing to 4th grade without the skills to cope. The reform was conceived to improve skills that educators viewed as key inputs to students' success in later grades (French, 2019; Povich, 2023). This anecdotal evidence suggests that around this time period, teachers and other school authorities ideally wished to only promote students who were ready for the next grade.

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, if promoted. A student is considered to demonstrate "any proficiency" if they receive a score of Level 2 or above according to the Department of Education guidelines in Table A1. We consider this binary outcome to be a proxy for whether a promoted student was ready for the 4th grade. Returning to the potential outcomes framework, this means that $Y_i(1) = 1$ if a student is ready for 4th grade and $Y_i(1) = 0$ if a student is not ready.[33]

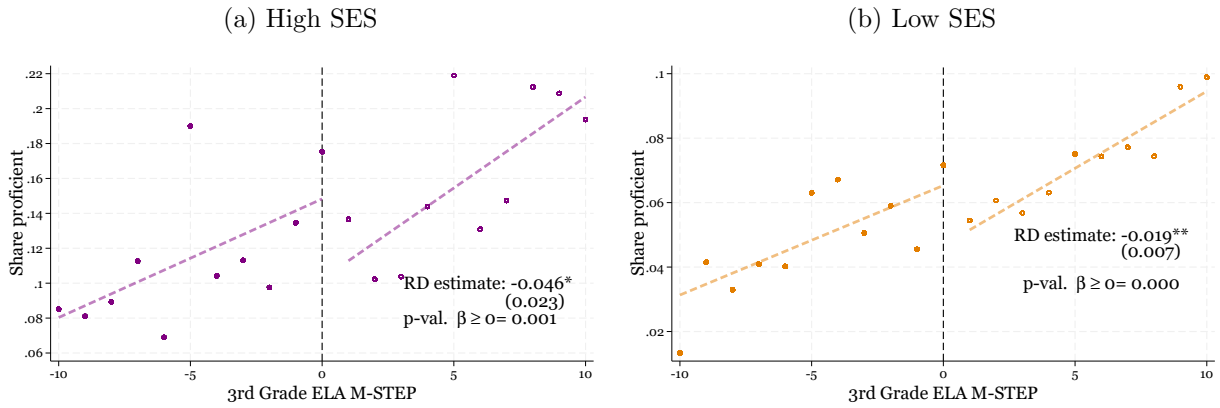
[31] Common exemptions include students who: are English language learners, have disabilities, and whose parents submit an exemption request. See Appendix A.1 for additional details.

[32] These students will eventually reach the 4th grade, but their outcomes at that point will be a function of being retained as well as being older for their grade. Our focus is on the upcoming school year, and we do not observe outcomes in 4th grade in the upcoming school year for retained students.

[33] 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 likely to

As before, we use quasi-experimental variation in promoted outcomes to estimate how average promoted outcomes varies 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 3 plots the relationship between the 3rd grade ELA M-STEP score and 4th grade outcomes but limits the sample 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 4.6pp (33%) less likely to demonstrate 'any proficiency' in 4th grade than promoted students below the cut-off. Marginal low SES students are 1.9pp (32%) less likely to demonstrate 'any proficiency' in the 4th grade than promoted students below the cut-off.

Figure 3: 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 in Table A1. The x-axis represents the running variable, the 3rd grade ELA M-STEP, re-centred by the cut-off of 1252. The sample only includes students who took the 3rd grade ELA M-STEP for the first time between 2020 and 2022, scored within 10 points of the cut-off, and were promoted to 4th grade. 'RD estimate' presents β from Equation 9. 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.

In an RD setting, the proportions and outcomes of always takers, never takers and compliers are identified by the intercepts of the local linear lines of best fit at the cut-off. Here, the proportion of always takers is the y-intercept in Figure 2 as the line approaches the cut-off from below zero. The proportion of compliers is the difference between the y-intercept in Figure 2 as the line approaches the cut-off from above zero and the proportion of always takers. Never takers are the remaining share of the relevant population. Similarly, the treated outcomes of always takers is the y-intercept in Figure 3 as the line approaches the cut-off from below zero. The treated outcomes of always takers and compliers together is the y-intercept in Figure 3 as the line approaches the cut-off from above zero. As described in Section 3, this is the information we need to estimate the average promoted outcomes for compliers, and bound average outcomes for never takers. Combining the estimates of average promoted outcomes for always takers, compliers and never takers for each SES group, we have the information to bound the average promoted outcomes if everyone from both

come in the Level 1 and Level 2 region, and binarizing as we do preserves power.

SES groups were to be promoted. This identifies average promoted outcomes at the cut-off, and not necessarily away from it.

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 imprecise and large relative to the first stage.[34] To mitigate the effects of noise on our estimates of the average outcome if all students were promoted, we adjust the bounding approach described in Section 3 to only use the **sign** of the discontinuity, rather than the **magnitude**. Figure 3 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 Section 3 involves assuming that never takers' proficiency rate if promoted is bounded above by that of compliers and below by zero. 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.[35] 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. The p-values presented in each Panel of Figure 3 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.[36]

With these adjustments, Figure 4 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. 14.4–14.8% of high SES students would demonstrate 'any proficiency' if promoted to 4th grade, while this is true for only 6.1-6.5% of low SES students. Using a bootstrapped inference procedure described in Appendix C.5, we reject the null that these bounds overlap ($p = 0.002$). Given that there are meaningful cross-SES differences in the underlying readiness for 4th grade, the raw SES gap in promotion is likely biased.

We now estimate bounds on the SES differences in promotion rates that condition on the outcome if promoted. Since each of the average outcomes if promoted in Figure 4 is estimated only at the cut-off, our baseline estimates will only use the average outcomes of promoted students **at the cut-off** and the promotion rate of students **at the cut-off** when following Equation 3. Such estimates are informative of SES discrimination in promotion rates **at the cut-off**.

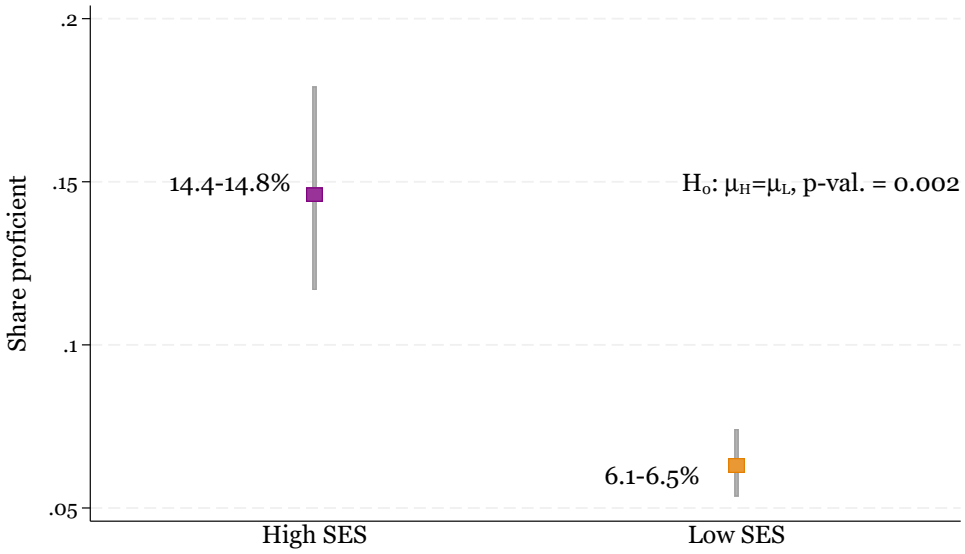
However, we can estimate discrimination below the cut-off (but within the window around the

[34] The discontinuity estimates might also be biased by other treatments potentially shifting around the cut-off. 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. As a result, students could be shifted between multiple treatments (not just promotion and retention) around the cut-off. Appendix A.1 discusses the conditions under which this would bias our analysis, and discusses existing evidence that suggests such bias is small.

[35] Figure A5 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 6.7% and 3% of the population of low and high SES students respectively.

[36] Figure A6 presents additional results validating the sign of the discontinuity in treated outcomes. We conduct a placebo exercise that re-estimates the RD specification in Figure 3, using every other possible test score in our analysis sample as the cut-off, while keeping the bandwidth the same. Each panel displays the distribution of resulting placebo RD estimates, and the vertical lines plot the observed estimates shown in Figure 3. Only 3.1% and 4.2% of placebo estimates for the high and low SES sample respectively are larger than the ones we observe.

Figure 4: Average outcomes if promoted, $E[Y_i(1)|R_i = r]$



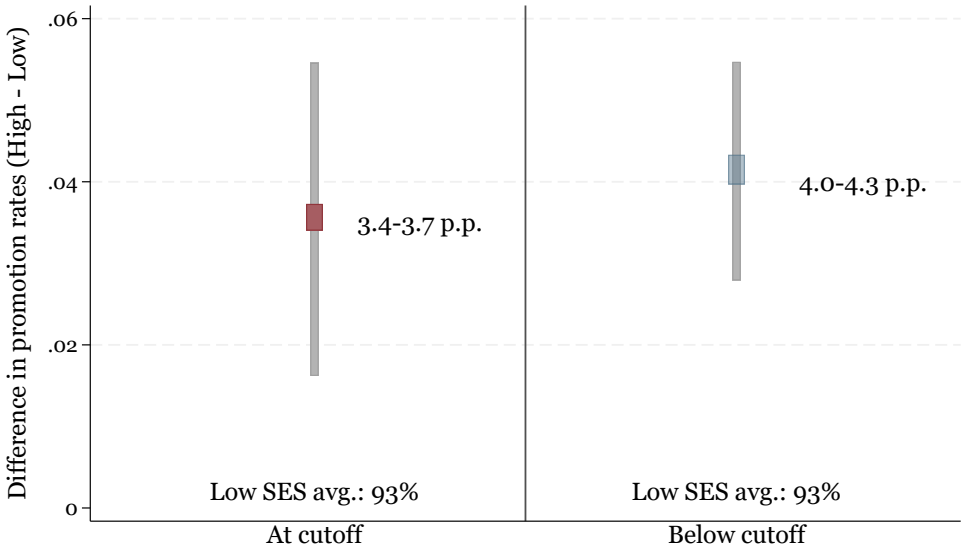
Note: This figure presents bounds on the average treated outcome obtained using the approach described in Section 3. The treatment is promotion and the treated outcome, $Y_i(1)$, is whether a student demonstrated any proficiency on both the Math and ELA M-STEP in 4th grade, as defined in Table A1. Confidence 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 C.5.

cut-off) under two additional assumptions. First, we need to assume that the running variable alone does not influence promotion rates. Columns 1 and 3 of Table A2 test this, failing to reject that promotion status varies with 3rd grade ELA M-STEP scores for either SES group within the small window around the cut-off. This provides suggestive evidence for the first assumption. Second, we need to assume that the average outcome if everyone at the cut-off were promoted is the same as the average outcome if everyone above or below the cut-off were promoted. This assumption is less likely to be satisfied—Columns 2 and 4 of Table A2 show that promoted outcomes clearly increase with the running variable.[37] The sign of the resulting bias on the discrimination estimands is ambiguous, as discussed in detail in Appendix C.6.

These assumptions are generally stronger than those described in Section 3 to estimate discrimination using difference-in-difference designs. In the difference-in-difference case, the assumptions restrict the impact of **time** on treatment and potential outcomes. Here, the assumptions restrict the impact of the **running variable** on treatment and potential outcomes. The running variable is likely to have a tight relationship with both of these factors in many settings, making such assumptions typically unlikely to hold. Given these issues, we focus on estimates of SES discrimination in promotion at the cut-off, but present estimates for students below the cut-off as well since it is arguably more policy-relevant than evidence only applicable at the test score cut-off.

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

Figure 5: 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 3. 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 in Table A1. Estimates below the cut-off are computed by applying estimates of the average treated from the cut-off to outcomes to all students below the cut-off. Confidence intervals are for the true parameter and are bootstrapped using 1,000 replications and a Bayesian bootstrap.

Figure 5 presents our estimates of the high–low SES differences in promotion rates under the different assumptions. The first bound (in maroon) shows that at the cut-off, high SES students are 3.4–3.7 p.p. (3.7–4%) 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. The next bound (in blue) provides a broader measure of discrimination, under the assumptions described above. High SES students below the cut-off are 4.0–4.3 p.p. (4.3–4.6%) more likely to be promoted than low SES students, even after accounting for unobservable differences.[38,39] Our 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).

Overall, 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. Our analysis also suggests that these gaps are concentrated among students who

[38] We rule out that these gaps are driven by SES differences in how likely parents are to request retention exemptions for their children (see Appendix A.1 for details on exemptions). We observe similar patterns in SES differences in average promoted outcomes (Figure A7) and SES differences in promotion rates that condition on these differences (Figure A8) if we exclude students whose parents requested that they be promoted instead of being retained.

[39] Disaggregating gaps from Figure 5 into promotion rates for students at the cut-off who would and would not be ready for 4th grade if promoted suggests that the SES gaps we find are driven by students who are not ready for the 4th grade (see Figure A9). Figure A9 also suggests that promotion rates for students who would be ready for 4th grade are weakly higher than that for students who would not be, consistent with school decision-makers targeting promotion towards students who are ready for the 4th grade.

are not ready for 4th grade. The estimates of SES disparities in promotion rates that we present here differ from alternative approaches. 'Selection-on-observables' approaches that control for gender, special education status, English language learner status, race, and district fixed effects would estimate high–low SES promotion gaps of 2.7 p.p., which is 22–28% smaller than the one that we find.

5 Racial discrimination in misdemeanor prosecution

In this section we use the approach described in Section 3 to measure racial discrimination in the decision to prosecute misdemeanor defendants. This is an important decision that can have adverse impacts on individuals' lives (Leasure, 2019; Mueller-Smith and Schnepel, 2021; Agan, Doleac, and Harvey, 2023). We find evidence that average outcomes if prosecuted differ by race, using difference-in-difference variation to isolate the quasi-experimental effects of a cut to the county prosecutors' budget in King County, Washington. Accounting for these racial differences in potential outcomes, we find no evidence of discrimination in prosecution before the reform. Even though prosecution rates fall due to the reform, we find that white defendants were more likely to be prosecuted than minority defendants after the reform. We find suggestive evidence that this is driven by prosecutors dropping low quality cases, which are more likely to be cases of minority defendants.

5.1 Natural experiment: King County budget reform

We study racial discrimination in misdemeanor prosecution in King County, Washington (Seattle metropolitan area and suburban areas) using administrative records on all criminal cases from 2000–2022 from the Washington Administrative Office of the Courts. We consider an individual as having their case prosecuted if their case **did not** meet the following condition: dismissed without requiring any punishments.[40] Our primary definition of a punishment excludes fines, but we assess robustness to including them. Our sample consists only of individuals in the court records, and so all of our estimates are representative of individuals who have been arrested and whose cases have been accepted by the prosecutors' office. This is not representative of the larger group of individuals who have contact with the Washington criminal legal systems.

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 the Western United States (Hu and Esthappan, 2017; Malott, 2024). Nevertheless, we later demonstrate that our results are robust to comparing prosecution rates between white (non-Hispanic) versus Black and Hispanic defendants.

[40] 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'.

As discussed in Section 3, 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 in September 2010, King County announced a large and unanticipated cut to the Prosecutor's Office budget. Facing a $60 million budget shortfall, the County cut the Prosecutors' budget by approximately $3.9 million, the equivalent of 33 full-time employees (Constantine, 2010). The Prosecutors' Office warned that this would reduce their ability to prosecute challenging and time-consuming cases, and that they would have to focus resources on offences they deemed to be high-priority (Ervin, 2010). The County tried to mitigate the budget cuts' impact on the criminal legal system by raising additional funds through a referendum to increase the sales tax. The referendum failed in November 2010, consigning the King County Prosecutors' Office to the new budget realities (Ballotpedia, 2010).[41]

The circumstances created by the budget reform present an opportunity to use our 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 the population into always takers, compliers and never takers for prosecution. We would then examine how outcomes if prosecuted vary across these groups, and then account for any differences.

We isolate the quasi-experimental variation using a difference-in-difference strategy that compares changes in prosecution and recidivism rates around the budget reform in King County, relative to changes in the adjacent counties unaffected by the reform.[42] We construct our analysis sample using criminal cases disposed of in the District Courts of these counties. District Courts are one type of court in which the County prosecutors work, and these courts typically hear criminal misdemeanor cases of varying severity. Given the messaging from the Prosecutors' Office regarding the types of cases they will find it difficult to pursue, cases in District Courts are most likely to be affected by the prosecutorial budget cuts. We limit to misdemeanor cases disposed of in the relevant District Courts between October 2008 and September 2012, a two-year span on either side of the budget cut announcement.[43] We construct re-offence outcomes and measure criminal history using information on cases filed by law enforcement in any Washington courts, including felony charges (filed in Superior Courts). These variables are not limited to the October 2008 and September 2012 interval.

The final sample, described in Table 2, consists of around 120,000 unique cases. 30% of the

[41] We can 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. Since the City Attorney has jurisdiction over Seattle's local municipal courts (and county prosecutors do not) local municipal courts are excluded from our sample. Hence, direct impacts of the City Attorney change are unlikely to be present in our analysis. Since this change occurred before this county budget cut, impacts 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 strong evidence of differential pre-trends during this period in our event studies examining prosecution, recidivism, and caseload composition (discussed below).

[42] The adjacent counties include: Snohomish, Pierce, Kitsap, Kittitas and Chelan.

[43] 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 is not a function of prosecution decision.

sample consists of non-white defendants, and this group is quite diverse—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 1 conviction in the past (averaging 3.5 prior convictions). Prosecution rates are quite high, but individuals are unlikely to face jail time if they prosecuted. Only 9% of individuals prosecuted in King County (and 6% overall) in our sample served any sentenced jail time. Among those who serve any time in jail, the average sentence is around 40 days long. By statute, the longest jail sentence for misdemeanors in Washington is one year. This occurs in only 6% of cases with some jail time, which represents 0.5% of prosecuted cases.

Table 2: Characteristics of Washington District Court sample

	Overall	King County	Adjacent Counties
N	122,156	51,242	70,914
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
Sentence length (Days) | Jail sentence	40.4	39.6	41.3

Note: Sample includes all criminal cases disposed of 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.

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 10, 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 this separately for white and minority defendants to ensure that we have enough variation in prosecution rates in each racial subgroup.

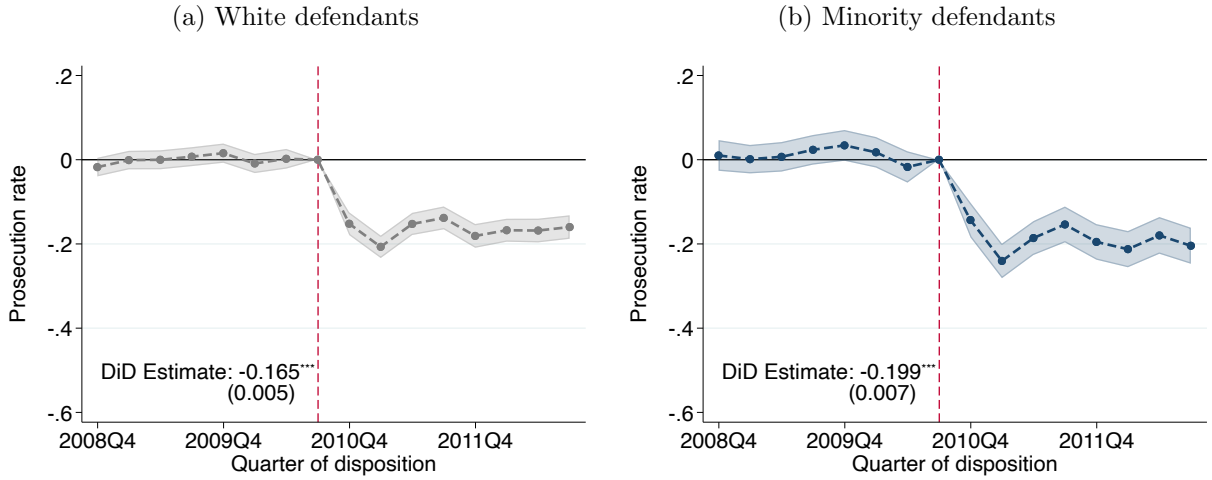
Figure 6 documents large drops in prosecution rates due to the King County budget reforms. Prosecution rates fall by 16.5pp and 19.9pp for white and minority defendants respectively (17.8% and 21.7% of pre-reform average prosecution rates in King County).[44] We do not see any evidence

[44] Figure B2 shows that the reduction in prosecution due to the reform is robust to using a more expansive definition of prosecution that considers cases listed as 'Dismissed' but with fines or fees as "prosecution". Using that definition, we see that prosecution falls by 15.9 p.p. and 18.8 p.p. for white and minority defendants, which is very similar to

of pre-trends in prosecution rates, which provides some evidence that this natural experiment is credible (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 shifts in prosecution suggest that we will have to bound outcomes for a relatively small share of the population.

$$D_{itg} = \theta_t + \delta \, \mathbb{I}[\text{King County}] + \sum_{k \neq -1} \beta_k \left(\mathbb{I}[t - \text{Budget Reform} = k] \times \mathbb{I}[\text{King County}] \right) + \varepsilon_{itg} \quad (10)$$

Figure 6: 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 2. '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 of on or after September 28, 2010, when the budget reform was announced. 95% confidence intervals are constructed using heteroscedasticity-robust standard errors.

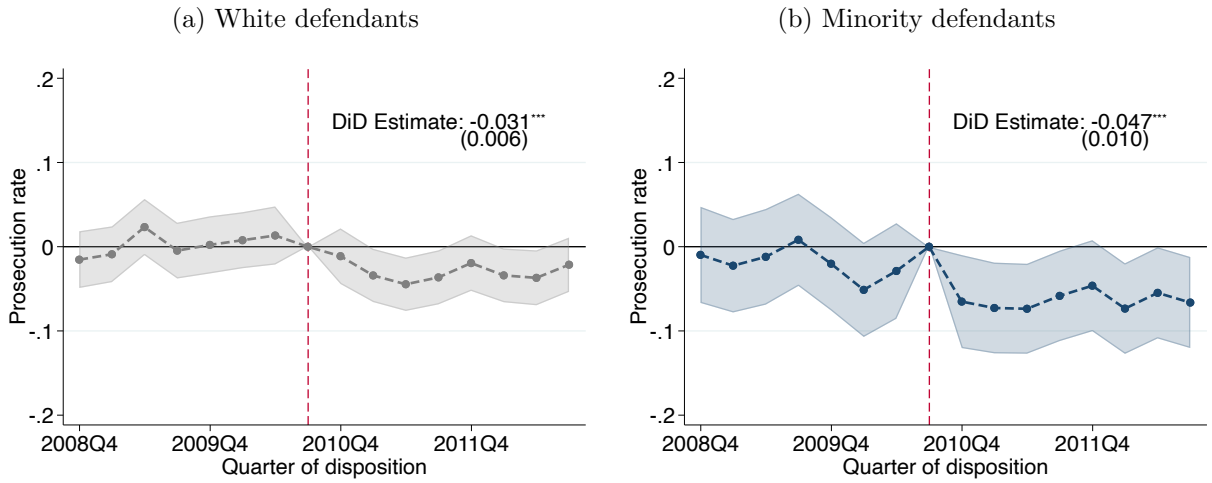
We next examine whether these shifts in prosecution rates influenced individuals' outcomes, which will reduce the width of the bounds for never takers' prosecuted outcomes. Figure 7 repeats the event study approach, assessing the impact of the budget reform on the probability that a defendant is charged with a new offence one year after disposition. We see that not being prosecuted reduces one year re-offence rates by 3.1pp for white defendants and 4.7pp for minority defendants (13.1% and 15% of pre-reform average re–offence rates in King County).[45] Again, we see no evidence of pre-trends in re-offence rates. The direction and magnitude of these estimates are consistent with recent work demonstrating that avoiding prosecution reduces future criminal activity (Mueller-

what we see with our main definition.

[45] These reductions are unlikely to be driven by incapacitation, given that only 9% of prosecuted defendants in King County served jail sentences. Among those who served jail sentences, the average length was 40 days, much shorter than the one year horizon used to compute the outcome (see Table 2). Finally, as discussed above, even though the maximum possible jail sentence for a misdemeanor in Washington is one year, this occurs for 0.5% of prosecuted cases in our sample. As a result, potential outcomes will not be mechanically censored by jail spells.

Smith and Schnepel, 2021; Agan, Doleac, and Harvey, 2023).[46,47]

Figure 7: 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 re-offence outcome includes any misdemeanor or felony charges filed against an individual anywhere in Washington State. Sample includes all misdemeanor defendants as described in Table 2. '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 of on or after September 28, 2010, when the budget reform was announced. 95% confidence intervals are constructed using heteroscedasticity-robust standard errors.

Next, we rule out different threats to the natural experiment's validity. We should be concerned if there are any 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, 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.

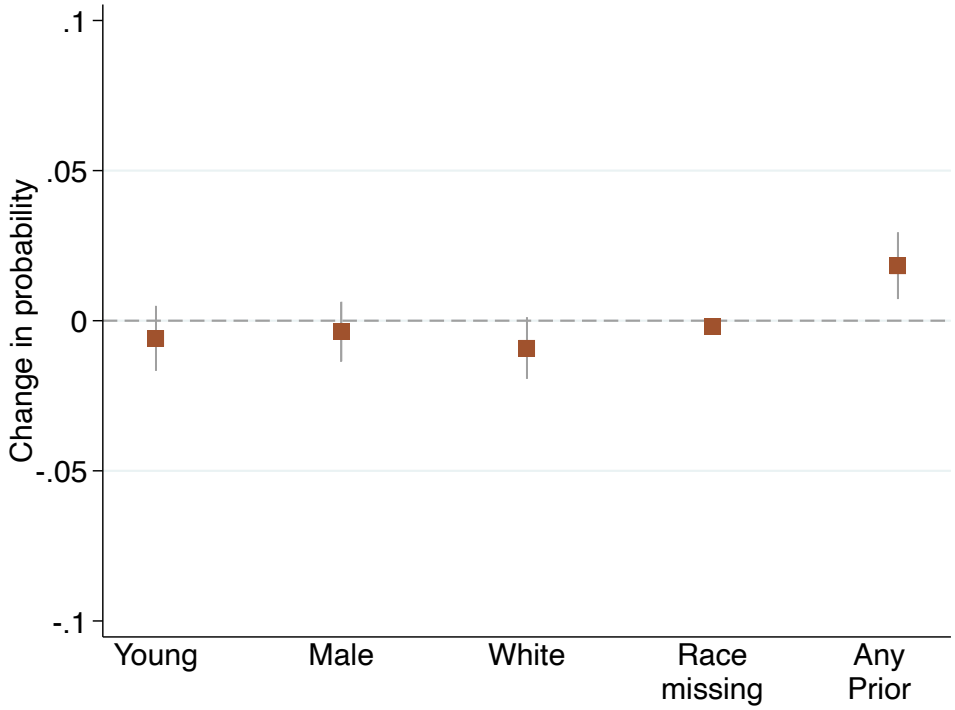
First, we test whether the composition of cases that prosecutors choose to accept from law enforcement changes discontinuously when the budget reform was enacted. We estimate a series of DiD regressions, where we compare the change in the share of individuals with a given baseline covariate before versus after the reform in King County to that same change in the adjacent counties. Figure 8 presents each of these coefficients and shows no evidence of compositional shifts in terms of most baseline characteristics (Figure B1 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 in

[46]These impacts are stable over time and persist even after two years, see Figure B3

[47]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 B4 presents results from an alternative specification that excludes 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 this contains information about charges that the prosecutor's office considers low-priority, we designate other charges as 'high-priority'. We see reductions in the probability of being charged with a new 'high-priority' offence that are similar to our baseline estimates in terms of their proportion of the relevant pre-reform means.

composition (3.5% of the pre-period mean in King County).[48]

Figure 8: 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.

We also test 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. Figure B5 & Figure B6 show no significant reductions in either officer employment or different categories of arrests. Figure B7 similarly finds no differential changes in house prices, unemployment rates or population, suggesting that the reform did not impact other economic factors that might influence criminal behaviour.

The results of these exercises suggest that the drop in prosecution rates and resulting reduction in recidivism are driven by the unanticipated budget reform, and not by concurrent policy or behavioral changes. While the King County budget reform is a valid natural experiment, we need to make the adjustments described in Section 3 to accommodate the DiD variation and use the

[48]This minor compositional shift might be due to the end of the federal investigation of Seattle Police Department (SPD) that resulted in lower SPD stops and arrests (Campbell, 2023). This could have contributed to this compositional shift if police focused on more serious offences. 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 B1). Our first stage and reduced form estimates are nearly identical if we exclude this time period from our sample (see Figure B8), and so we proceed with the full sample for our baseline analysis.

reform to estimate discrimination. Next, we describe the adjustment and associated assumptions in the context of this setting. We then use this natural experiment to estimate racial differences in prosecution among individuals who would have the same outcomes if prosecuted.

5.2 Estimating discrimination in prosecution using the budget reform

We begin by mapping the objects in the empirical discussion to the potential outcomes framework discussed in Section 3. Individuals are considered 'treated' if they are prosecuted ($D_i = 1$) and considered 'untreated' if dismissed. Potential outcomes in each treatment state are the binary indicators of recidivism used in the event study estimates in Figure 7. That is, $Y_{it}(d) = 1$ if an individual assigned to treatment state $D_i = d$ would be charged with a new offence in Washington State within one year of disposition. Z indicates periods around the King County budget reform. Next, we validate the adjustment requires to use the budget reform DiD to estimate discrimination, and then apply our approach to estimate racial discrimination in prosecution among individuals who would have the same re-offence outcome if prosecuted. We then discuss estimating discrimination conditional on having the same re-offence outcome if dismissed.

5.2.1 Empirically validating the difference-in-difference adjustments

As described in Section 3, 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 race were prosecuted. This would help us understand whether the unobserved potential outcomes vary by race, in which case we adjust the discrimination estimates for them. We use the trend in outcomes among prosecuted individuals adjacent counties as an estimate of the time trend that prosecuted individuals in King County would have experienced had the budget reform not occurred. Using this, we purge the change we observe in King County of the change only due to time.

In this context, we can estimate discrimination using the DiD if we assume the following conditions hold within each race group:

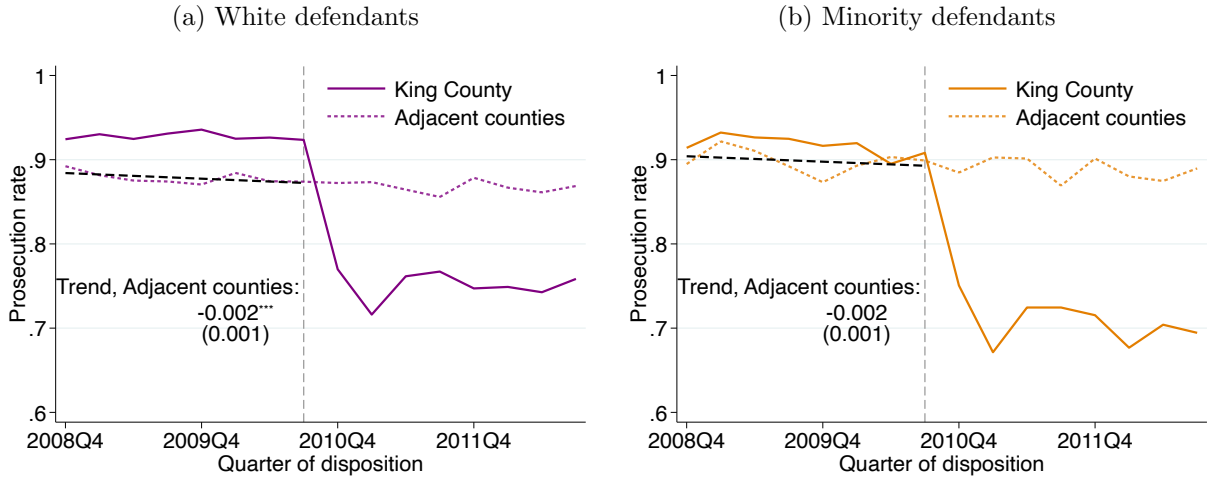
A1 Time does not shift individuals' prosecution status.

A2 Re-offence outcomes if prosecuted, $Y_{it}(1)$, would trend similarly for always takers & compliers and is independent of county absent the budget reform.

While these assumptions are fundamentally untestable, we provide evidence that they are not grossly violated here. Figure 9 examines **A1**, testing for time trends in prosecution using pre-period data from the adjacent counties, separately for white and minority defendants. These shifts in prosecution rates are very small. Even though estimated trend for white defendants is significant, it represents a 0.2 p.p. reduction in prosecution rates every quarter, which is approximately 1% of the reductions in prosecution that we see in Figure 6. To ensure that these aggregate trends are not masking shifts into and out of prosecution for different types of people, we repeat this exercise separately by covariate subgroups (gender, criminal history, age). Figure B23 shows that trends in

prosecution rates are similarly very small even within these subgroups. This builds confidence that individuals are not shifting into or out of treatment over time.

Figure 9: Testing trends in prosecution rates in adjacent counties



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

Next, we test **A2**, which requires parallel trends to hold between always takers and compliers. This is because we need to account for time trends in the average prosecuted outcomes for compliers and always takers so we can use that information to construct bounds for never takers. Since we cannot identify always takers and compliers in the adjacent counties, we test for differential pre-trends by subgroup (gender, criminal history, age), which might be correlated with being an always taker or complier. Using pre-period data, Figure B24 shows that there is no evidence of differential trends in $Y_{it}(1)$ across counties within the various subgroups. Figure B25 conducts a similar exercise, finding limited evidence of differential trends in $Y_{it}(1)$ within counties but across subgroups. While this is not definitive evidence that these assumptions are satisfied, they build credibility that they are reasonable in this setting.

5.2.2 Baseline estimates of discrimination

We now use the variation from the budget reform DiD to 1) estimate if there are racial differences in the average re-offence outcome that we would see if all defendants of each race were prosecuted and 2) account for any such differences. We denote Z as indicating periods around the King County budget reform. Since Y_{it} varies with time, we estimate average prosecuted outcomes in each period t: $E[Y_{it}(1)|R_i = r]$. Figure B9 presents event study estimates showing how the reform impacts prosecuted outcomes – we use this shift in prosecuted outcomes to estimate average prosecuted outcomes for always takers and compliers.

Figure 10 displays how average outcomes if prosecuted vary across always takers, compliers, and never takers for white and minority defendants using the IV approach and weak monotonicity

assumptions discussed in Section 3.[49] Panels a) and c) present estimates from before the reform, and Panels b) and d) present estimates from after the reform. White defendants' average outcomes for always takers and compliers are uniformly lower than the corresponding averages for minority defendants. This suggests that there likely are racial differences in the average re-offence outcomes that we would see if all defendants were prosecuted. Additionally, we see that never takers, comprise 7-8% of the population of each race group. This implies that 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 are likely to be relatively tight. Since the bounds on the average outcomes if everyone were prosecuted are inputs into estimating the discrimination estimands (see Equation 3), we should expect relatively tight bounds on discrimination as well.

As discussed earlier, assuming that the average potential outcomes if prosecuted is weakly monotonic across always takers, compliers, and never takers implicitly places assumptions on the underlying decision-maker behavior. Here, we assume that the defendants whom prosecutors are least likely to prosecute (never takers) are at least as likely to re-offend if prosecuted as marginal defendants. This assumption might be violated if other inputs into prosecution decisions are correlated with re-offence outcomes in specific ways. For example, this assumption would be violated if i) never takers are not prosecuted since their cases are backed up by low quality evidence, and ii) if individuals with low quality cases are systematically unlikely to re-offend if prosecuted. While we do not have reason to believe that such violations occur in our setting, we discuss robustness to bounds that do not assume weak monotonicity below, and find similar results.

Figure 11 estimates bounds on the average prosecuted outcome by computing a weighted average of the average outcomes for always takers, compliers, and never takers from Figure 10. We see meaningful and significant cross-race differences in the average outcomes that would be realized if all defendants were prosecuted. Before the budget reform, 25–29% of white defendants would have re-offended if prosecuted, while 32–37% of minority defendants would have done so. Using a bootstrapped inference procedure described in Appendix C.5, we reject the null that these bounds overlap ($p = 0.004$). After the reform, the bounds on average prosecuted outcomes are still meaningfully different (20–25% vs 27–32%), although testing the probability that they overlap is less precise ($p = 0.105$).[50] Given that outcomes if prosecuted differ meaningfully by racial group, not accounting for unobservable cross-race differences would yield incorrect estimates of discrimination.

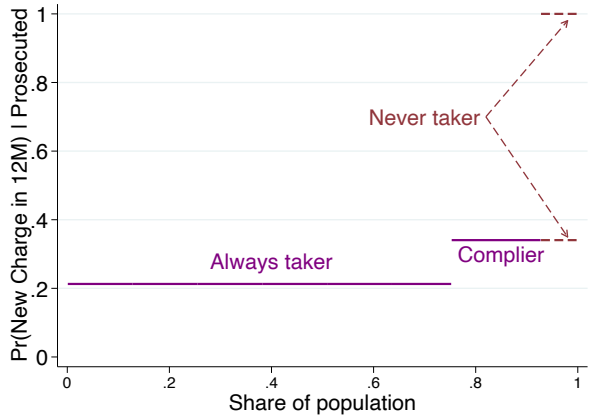
Next, we estimate bounds on the racial differences in prosecution rates that condition on outcomes if prosecuted. Following Equation 3, we need, for each race, 1) the average re-offence

[49] 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 large and significant reductions in prosecution similar to the estimates in Figure 6 across all race × covariate cells, suggesting that 'defiers' are unlikely to be present here (Figure B10).

[50] These patterns are not due to our decision to define a broad minority subsample. Figure B11 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.

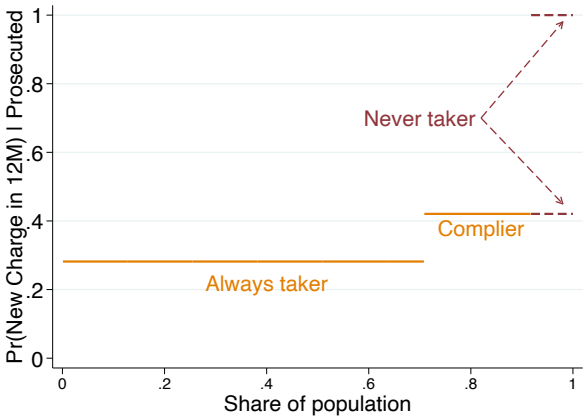
Figure 10: Average outcomes if prosecuted ($Y_i(1)$) by compliance group

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



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

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



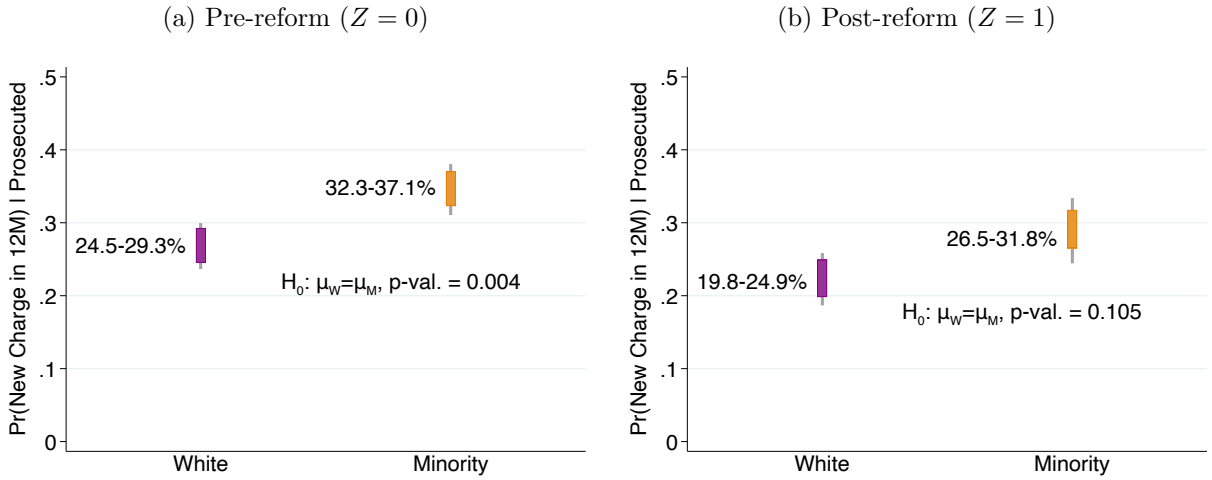
(d) Minority defendants, post-reform ($Z = 1$)

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 is charged with a new offence within 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 $Y_i(1) \in \{0, 1\}$.

outcomes among prosecuted defendants in each period, 2) the prosecution rate in each period 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 11.

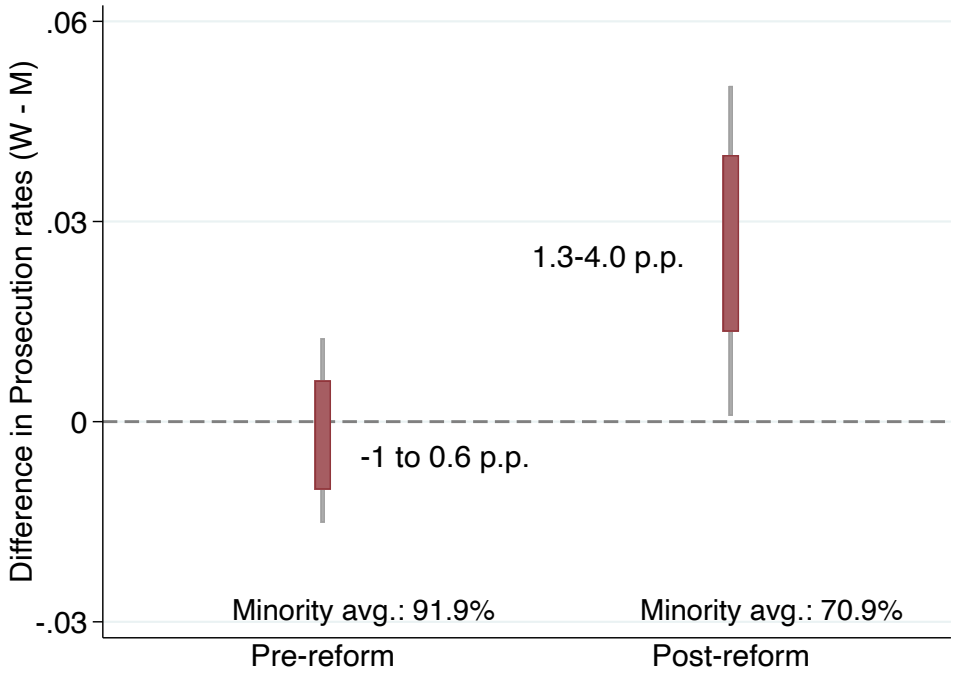
Figure 12 displays bounds on the average white–minority gap in prosecution rates that is conditional on outcomes if prosecuted. After accounting for racial differences in the outcomes if prosecuted, we cannot reject that white and minority defendants were prosecuted at similar rates before the budget reform. Our bounds suggest that white defendants before the reform were between 0.6 p.p. **more likely** to 1 p.p. **less likely** to be prosecuted compared to their minority counterparts. While this suggests that prosecution in this context may not have been discriminatory before the reform, this does not rule out discrimination in other stages in the criminal legal system or other aspects of society.

Figure 11: Average 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 3, separately by race and time period. The treatment is prosecution and the treated outcome, $Y_i(1)$, is whether an individual is charged with a new offence within 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 C.5.

Figure 12: 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 3. $Y_i(1)$ is whether an individual is charged with a new offence within one year after disposition, if prosecuted. Confidence intervals are for the true parameter and are bootstrapped using 1,000 replications and a Bayesian bootstrap.

This pattern changes dramatically after the reform—white defendants were 1.3–4 p.p. (1.8–5.6%) more likely than minority defendants to be prosecuted, after accounting for racial differences in the outcomes if prosecuted. That is, even though the budget reform results in prosecution rates falling overall, they fall by a greater amount for minority defendants than for white defendants, even after conditioning on outcomes if prosecuted.

These findings are robust to changing our empirical definitions of racial groups, prosecution, and re-offence outcomes. First we show that these patterns are not driven by our definition of the 'minority' group. Figure B12 and Figure B13 present the average re-offence outcomes if prosecuted and discrimination estimates that include only Black and Hispanic defendants in the minority group. These estimates are qualitatively and quantitatively similar to our baseline. We also show that our findings are robust to a more expansive definition of prosecution that includes fine-only punishments. Figure B14 shows racial gaps that are qualitatively similar and contained within the baseline bounds. The racial gaps in prosecution that we estimate remain stable even if we reduce or expand the time horizon used to measure re-offences (see Panel (a) of Figure B15).

Our results are also robust to weakening key identifying assumptions. The patterns observed in Figure 12 are not driven by the weak monotonicity assumption imposed in Figure 10. Figure B16 shows that relaxing the weak monotonicity assumption by allowing never takers' outcomes to lie between 0 and 1 yield nearly identical discrimination estimates as in our baseline.

Finally, Figure B17 displays patterns similar to our baseline when we simultaneously adjust for potential re-offence outcomes and baseline covariates (age, gender, criminal history, court fixed effects). This implies that covariates in this context do not provide additional information that is not already captured by the potential re-offence outcomes. This also suggests that the unwarranted disparities that we observe are not being mediated by the covariates that we have access to.

Alternative approaches to estimate discrimination here yield estimates that are outside of the bounds that we estimate. 'Selection-on-observable' estimates that control for age, gender, criminal history, and court fixed effects would estimate white–minority prosecution gaps of 1.2 p.p. before the budget reform and 4.4.p.p. after the reform. This suggests that alternative approaches to estimate of discrimination in prosecution before the reform might be incorrectly signed, and that estimates of discrimination after the reform would be biased by 10%–238%.

5.3 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 overall prosecution rates fell after the reform, white defendants were more likely to be prosecuted than minority defendants who would experience identical outcomes if prosecuted.[51] Next, we investigate potential factors that

[51] Appendix B.3 examines how prosecution rates vary by functions of potential outcomes. We see suggestive evidence that the post-reform racial gaps are more pronounced among the set of individuals who would not re-offend if prosecuted. We also find that prosecution is less likely for defendants who would commit a new offence if prosecuted. Under additional assumptions restricting the evolution of dismissed outcomes over time and restricting the support of treatment effects of prosecution, we also find suggestive evidence that prosecution is less likely in cases where it

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 be due to cases involving minority defendants requiring more resources to prosecute. Discrimination in previous stages of the criminal legal system, such as policing (Goncalves and Mello, 2021; Owens and Ba, 2021; Jordan, 2024), could also lead cases involving minority defendants to be backed up by weaker evidence. Since cases with weak evidence require greater resources to prosecute successfully, these cases are less likely to be pursued after the budget reform when resources became scarce. Instead, prosecutors might shift their resources to cases that they were more likely to win. As described above, statements from the King County Prosecutors' Office suggested they were concerned about not being able to prosecute resource-intensive and time-consuming cases.

We investigate this explanation by repeating our approach to estimate discrimination in two subsamples that may vary in terms of average case quality. We classify offence categories into two bins based on the share of charges that end up being successfully punished in King County, using (pre-reform) data from September 2004 to September 2010. The logic is that for offences where the average case is typically high quality, we should see a high conversion rate of charges into punishments. Offences for which a relatively high share of charges end up being punished are: drug, DUI, property, prostitution and weapons violations. We refer to these as "high quality" cases. On the other hand, arrests for traffic, 'other' and violent offences have a relatively low share of charges that are punished, and we refer to these types of cases as "low quality".[52,53]

Figure 13 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 1.6–4.6 p.p.more likely than minority defendants to be prosecuted, after accounting for racial differences in the outcomes if prosecuted. In contrast, we see a gap of only 0.2 to 1.1 p.p. among the low quality subset of cases, which is much smaller and lies outside of the post-period high quality bounds. However, we cannot reject that these bounds overlap due to the reduction in power from splitting our data into two subsamples and estimating discrimination within them.[54,55]

While it is only suggestive evidence, the divergence in discrimination patterns between the "High Quality" and "Low Quality" subsamples suggests that minority defendants' cases are lower quality, and that such cases are dropped due to the budget reform. Taken seriously, this implies

would be harmful, i.e., by generating new crime.

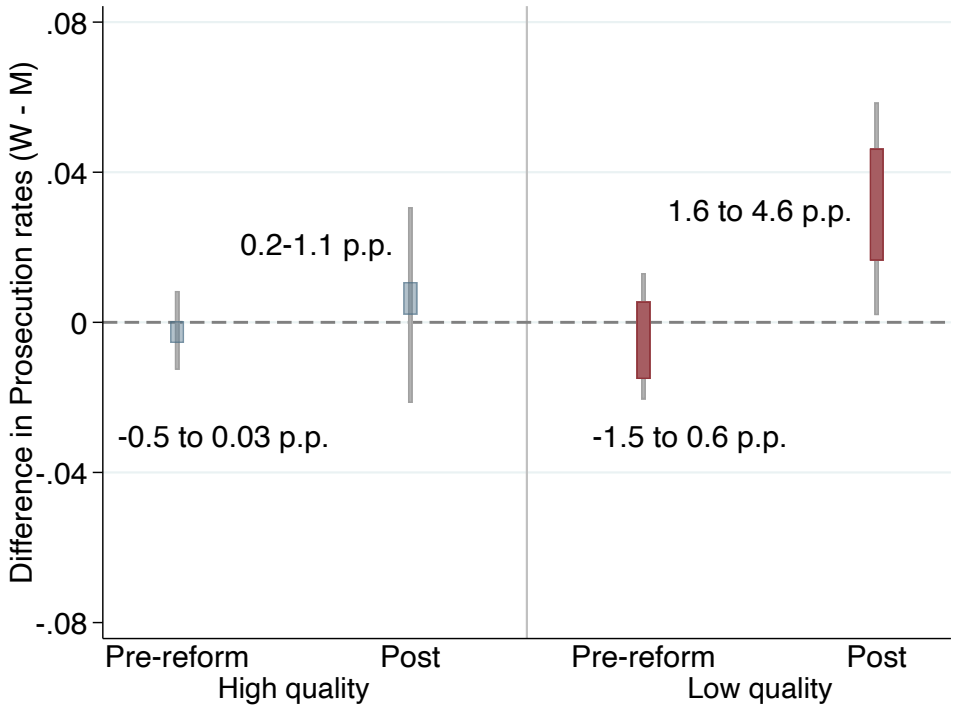
[52]Admittedly, other factors may vary between these two categories, e.g., resource intensity.

[53]Some cases in the data are dismissed without charges ever being recorded. We consider these cases as 'low quality', but the results discussed below are robust to excluding such cases from the exercise.

[54]An alternative exercise is to define the prosecuted outcome as whether the case was successfully sentenced to any punishment, and condition on that. However, such gaps would be biased by the exclusion of potential re-offence outcomes. On the other hand, the exercises above on re-offence outcomes as well as an admittedly imperfect measure of case quality. Nevertheless, Figure B18 conditions on the potential punishment outcome if a case was prosecuted. Racial gaps from this exercise are similar to the baseline analysis in Figure 12, and we cannot reject that the bounds between the two cases do not overlap.

[55]These findings are qualitatively similar but less precise if we exclude cases with missing charge information (Figure B19).

Figure 13: 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 3. $Y_i(1)$ is whether an individual is charged with a new offence within 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.

that prosecutors were pursuing most cases before the reform, and passing through any pre-existing disparities from prior stages in the criminal legal system (Harrington and Shaffer, 2024). However, in shifting their focus to high quality cases, prosecutors may have been undoing disparities from prior stages in the criminal legal system. This behavior contrasts with the behavior of other agents when fiscally-constrained. For example, police alter their search behavior in ways that increase disparities when budget deficits bind (Makowsky, Stratmann, and Tabarrok, 2019). However, the patterns that we find are consistent with prior work showing how prosecutors can use their discretion to attenuate discrimination from pre-prosecution decisions (Harrington and Shaffer, 2023).

5.4 Conditioning on outcome if dismissed, $\mathbf{Y_i(0)}$

So far, we have defined discrimination as racial differences in prosecution, conditional on the outcome if prosecuted. We consider an alternate definition of discrimination that conditions on 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 Section 3 (Equation 7), this 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 10), we should expect the bounds on average outcomes and discrimination to be wider here.

Again, given the time-varying nature of the potential outcomes, we need to 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. We make a similar parallel trends assumption as we previously did, but now for the re-offence outcomes if **dismissed**. As described in **A3**, we require parallel trends to hold between never takers and compliers (and not for always takers) because we need to account for time trends in the average dismissed outcomes for compliers and never takers, and construct bounds for always takers' dismissed outcomes.[56]

A3. Re-offence outcomes if dismissed, $Y_{it}(0)$, would trend similarly for never takers & compliers and is independent of county absent the budget reform.

Using this assumption, we isolate the shift in re-offence outcomes if dismissed that is due to the budget reform. Figure B20 presents estimates of the average re-offence rates if all individuals were dismissed, separately by race and time period (since potential outcomes if dismissed are allowed to vary with time). We see suggestions that minority defendants are more likely to commit a new offence if dismissed, but we cannot reject that the bounds do not overlap. For example, our estimates suggest that before the reform, between 4.3% and 15.6% of white defendants would commit a new offence if all were dismissed, while between 6.4% and 19.7% of minority defendants would commit a new offence if dismissed.[57]

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 B21 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 0.9–8.2 p.p. (1.3–11.6%) more likely to be prosecuted after the reform. These findings are robust to altering the time horizon used to measure re-offences (see Panel (b) of Figure B15) and also controlling for baseline covariates (see Panel (b) of Figure B17).

We 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. We see patterns that are qualitatively similar to what we observe when conditioning on re-offence outcomes if prosecuted, but less precise. In Figure B22, the white–minority gap after the reform lies in [1.4 p.p., 8.5 p.p.] for "low quality" cases and lies in [–1.4 p.p., 6.8 p.p.] for "high quality" cases. While these patterns are less clear-cut than in Figure 13, the bounds on post-period gaps for

[56] Similar to the previous validation exercises, we test for differential pre-trends in the outcomes of dismissed individuals across county and covariate subgroups. Figure B26 shows no evidence of differential pre-trends in re-offence outcomes if dismissed across counties but within various demographic subgroups. Figure B27 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.

[57] The similarity of these bounds across race contrasts to other work that finds minority defendants have higher risk of re-offending (Arnold, Dobbie, and Hull, 2022). One potential reason for our findings is discrimination in policing, which could push the race-specific re-offence distributions closer together if the marginal minority defendant arrested has a lower "risk" of re-offending. However, the bounds are too wide to prove this claim.

"low quality" cases do not contain zero, while the post-reform bounds among "high quality" cases do contain zero and are quite imprecise.

5.5 Summary

The results of our analysis represent 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 reduces overall prosecution rates, white defendants are more likely to be prosecuted than minority defendants with similar potential re-offence outcomes. Digging deeper, our evidence suggests that prosecutors seem to be dismissing low quality resource-intensive cases, attenuating discrimination in prior stages of the criminal legal system. Finally, our findings that not being prosecuted due to the budget reform reduces future criminal activity adds to the literature on the impacts of non-prosecution.

6 Conclusion

This paper shows how to use a natural experiment that generates a binary instrumental variable (IV) to estimate discrimination conditional on potential outcomes or treatment effect. We combine the shift in treatment decision rates induced by the binary IV with assumptions on the relationship between selection into treatment and average potential outcomes which are common in the marginal treatment effects literature. Depending on the strength of the assumptions, we obtain bounds or point estimates. Our approach measures discrimination when groups are unobservably different and individuals are not randomly assigned to decision-makers, but a natural experiment is available. Thus, we expand the set of places where researchers can study discrimination while accounting for unobserved differences in potential outcomes.

Our implementation with a difference-in-difference (DiD) strategy also adds to the DiD-IV literature. We show how to use selection into being an always taker, never taker or complier to estimate average potential outcomes among units that *experienced an intervention* using information on units who *did not experience the intervention*. We also present the relative advantages and disadvantages of using a regression discontinuity (RD) approach to estimate discrimination.

We study racial discrimination in misdemeanor prosecution, using a budget cut in King County, Washington that sharply reduced prosecution rates. Using a DiD strategy, we find significant racial differences in the unobserved re-offence outcomes. Adjusting for these differences, we find no evidence of discrimination in prosecution before the budget reform. After the reform, white defendants were more likely to be prosecuted than minority defendants with identical potential re-offence outcomes. 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.

We also study discrimination by socio-economic status (SES) in the decision to promote Michigan public school 3rd graders to 4th grade, using a test score RD design. We find significant SES differences in the underlying probability of succeeding in 4th grade in a window around a test score

cut-off. Even after accounting for these underlying differences, high SES students were more likely to be promoted than low SES students, suggesting that disparities documented by prior work in this context are not solely driven by unobserved 4th grade outcomes (Westall et al., 2022a,b).

While our analysis conditions on a single potential outcome at a time, future methodological steps might incorporate insights from work estimating average population outcomes while accounting for two dimensions of unobservable heterogeneity (Dutz et al., 2021). Such an approach could quantify discrimination between individuals similar on more than one dimension, e.g., discrimination in prosecution conditional on potential re-offence and employment outcomes.

Fruitful next steps for the empirical analysis could assess the mechanisms underlying how discrimination is affected by changes in discretion (in the case of student grade promotion) and resource constraints (in the case of prosecution). Focusing on the prosecutorial context, data that can track individual events within each court case would allow us to better understand which prosecutorial actions amplify versus offset any pre-existing disparities.

References

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

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

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, 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.

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.

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., 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.

Chalfin, Aaron, and Justin McCrary. 2017. "Criminal Deterrence: A Review of the Literature." *Journal of Economic Literature* 55 (1): 5–48.

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."

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.

Eren, Ozkan, Michael F. Lovenheim, and H. Naci Mocan. 2022. "The Effect of Grade Retention on Adult Crime: Evidence from a Test-Based Promotion Policy." *Journal of Labor Economics* 40 (2): 361–395.

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."

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, William Murdock III, and Hannah Shaffer. 2023. "Prediction Mistakes in the Criminal Justice System: Evidence from Linking Prosecutor Surveys to Court Records." Tech. rep.

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."

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."

Humphries, John Eric, Aurelie Ouss, Kamelia Stavreva, Megan T Stevenson, and Winnie van Dijk. 2023. "Conviction, Incarceration, and Recidivism: Understanding the Revolving Door." .

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.

Jacob, Brian A, and Lars Lefgren. 2009. "The Effect of Grade Retention on High School Completion." *American Economic Journal: Applied Economics* 1 (3): 33–58.

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

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.

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.

Locke, Victoria Nevin, and P. Johnelle Sparks. 2019. "Who Gets Held Back? An Analysis of Grade Retention Using Stratified Frailty Models." *Population Research and Policy Review* 38 (5): 695–731.

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.

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.

Moller, Stephanie, Elizabeth Stearns, Judith R. Blau, and Kenneth C. Land. 2006. "Smooth and rough roads to academic achievement: Retention and race/class disparities in high school." *Social Science Research* 35 (1): 157–180.

Mountjoy, Jack. 2022. "Community Colleges and Upward Mobility." *American Economic Review* 112 (8): 2580–2630.

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.

Paufler, Noelle A., and Audrey Amrein-Beardsley. 2014. "The Random Assignment of Students Into Elementary Classrooms: Implications for Value-Added Analyses and Interpretations." *American Educational Research Journal* 51 (2): 328–362.

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."

Reeves, James. 2023. "Discrimination and Permissible Considerations." In progress.

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." .

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.

Strunk, Katharine O., Tanya S. Wright, John Westall, Qiong Zhu, Tara Kilbride, Amy Cummings, Andrew Utter, and Madeline Mavrogordato. 2022. "Michigan's Read by Grade Three Law: Year Two Report." Tech. rep.

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, Katharine O Strunk, and Andrew Utter. 2023. "Challenges in Implementing Teacher-Student Assignment Policies: Evidence From Michigan's Read by Grade Three Law." .

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." .

Appendix A Additional results & details: Student grade promotion

A.1 'Read by Grade 3' policy details

We focus on the aspect of the 'Read by Grade 3' policy that required students who scored below 1252 (approximately the 5th percentile) on the Michigan standard reading test, the English Language Arts Michigan Student Test of Educational Progress (ELA M-STEP), to be retained. The exact mapping of test scores to the policy guidance is below:

- Score $\leq$ 1252: The policy mandated that these students were to be retained unless they had a 'good cause exemption', which is described further below. These students are also supposed to receive additional supports to improve their reading skills. The supports included extra instructional time and improved instructional quality over the next school year.

- Score in [1253,1271]: The policy did not mandate that these students were to be retained. These students were eligible for the additional supports that students who scored below 1252 were mandated to receive. However, this provision was not mandatory

- Score $\geq$ 1272: These students are to be promoted, and there is no direct mandate or encouragement for them to receive additional academic supports.

While students who scored $\leq$ 1252 on the ELA M-STEP were supposed to be retained, students who met certain criteria were exempt from this requirement. This included the following students: English language learners with $\leq$ 3 years of English instruction, students with disabilities (i.e., with a Section 504 Plan or an Individualized Education Plan), students who have been retained before and have received supports for $\geq$ 2 years, students who have been enrolled in the district for $\leq$ 2 years and were not previously designated as having reading issues, students who later demonstrate proficiency through alternate assessments and students whose parents submit an exemption request (subject to superintendent approval) (Westall et al., 2022b).

Given that the RBG3 policy mandate includes the provision of additional academic interventions, and that this may change around the 1252 cut-off, there is concern that moving from one side of the cut-off to another may shift students across multiple treatments, biasing our estimates of average outcomes if promoted. We investigate potential bias by decomposing the reduced form treatment effect of promotion identified by this design into margin-specific treatment effects using the logic of Humphries et al. (2023).

Define the following treatment states: retention & supports (rs), promotion & no supports (pn) and promotion & supports (ps).[58] Let $D_i = j$ if individual i received treatment j. For simplicity,

[58] We rule out retention & no supports since a survey of school principals conducted by Michigan State University researchers, described in more detail below, suggests that almost all of the retained students receive some additional intervention (Strunk et al., 2022; Berne et al., 2023).

consider the RBG3 test score cut-off rule as a binary instrument, where $Z = \mathbb{I}(\text{Score}>1252)$. We then have that $Pr(\text{Promoted}|Z = z) = Pr(D_i = pn \text{ or } D_i = ps|Z = z)$ is increasing in Z.

Adopting the conditional pairwise monotonicity (CPM) assumption from Humphries et al. (2023), we permit the following set of treatment flows that occur when moving from $Z = 0$ to $Z = 1$:

- Retention & supports to promotion & no supports: $rs \to pn$
- Retention & supports to promotion & supports: $rs \to ps$
- Promotion & supports to promotion & no supports (or the reverse): $ps \to pn$ (or $pn \to ps$)

This assumption ensures that flows between treatments are weakly one way and forces individuals to only flow out of retention, not into it.

Let $\omega_{p\to q}$ denote the proportion of individuals flowing in direction $p \to q$ and let $\triangle^{j-k}_{p\to q}$ denote the treatment effect between treatments j and k for people shifted along the p and q margin. Finally, let the third flow be $ps \to pn$ for simplicity.

We then rewrite the reduced form estimate from a regression of the form $Y_i = \alpha + \beta Z_i + \varepsilon_i$ as the following expression:

$$\beta = E[Y(Z = 1) - Y(Z = 0)] = \omega_{rs\to pn} \triangle^{pn-rs}_{rs\to pn} + \omega_{rs\to ps} \triangle^{ps-rs}_{rs\to ps} + \underbrace{\omega_{ps\to pn}\triangle^{pn-ps}_{ps\to pn}}_{\text{Bias}}$$

The above expression shows that the reduced form estimates of the effect of the RBG3 test score cut-off on outcomes is a weighted average of the treatment effect of 1) promotion without supports relative to retention ($\triangle^{pn-rs}$), 2) promotion & supports relative to retention ($\triangle^{ps-rs}$) and 3) additional supports with promotion relative to promotion without supports. A weighted average of the first two treatment effects can be interpreted as a total effect of promotion relative to retention and does not necessarily represent bias. However, term 3) contaminates the reduced form estimates with the effect of supports.[59]

This bias term will be irrelevant if $\omega_{ps\to pn} = 0$. While it is not possible to test this directly, we rely on analysis of a survey of around 300 principals conducted by researchers at Michigan State University (Strunk et al., 2022; Berne et al., 2023). Each principal was asked for a list of the additional supports they provided students in 3 different categories during the 2021-22 school year: students who scored below 1252 and were retained, students who scored below 1252 and were promoted, and students who scored above 1252. For each of the 11 possible interventions listed in the survey, they compare the difference between the share of principals who report offering the intervention to a) students who scored below 1252 and were promoted and to b) students who scored above 1252 and were promoted. The change in the probability of receiving any intervention

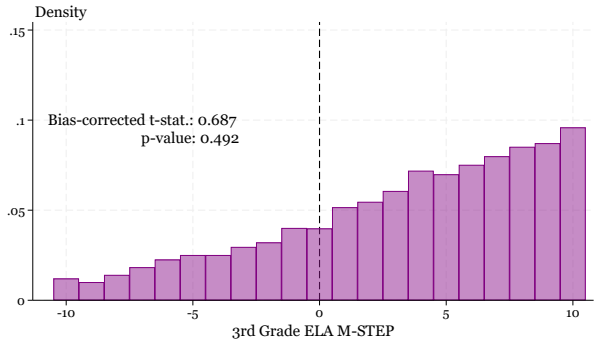
[59]This bias term arises under the CPM assumption, but not under more restrictive monotonicity assumptions. For example, if we were to assume unordered partial monotonicity instead, the only permissible flows would be $rs \to pn$ and $rs \to ps \implies \omega_{ps\to pn} = 0$, implying that the bias term is 0 (Mountjoy, 2022) .

around the cut-off is small (<5%) and is marginally significant ($p = 0.1$). Additionally, a recent study one of the key interventions using the RBG3 RD design finds no evidence of changes in the probability of being assigned a "highly effective" teacher around the test score cut-off (Westall, Strunk, and Utter, 2023). Taken together, we interpret these findings as evidence that this type of bias is minimal in this context.

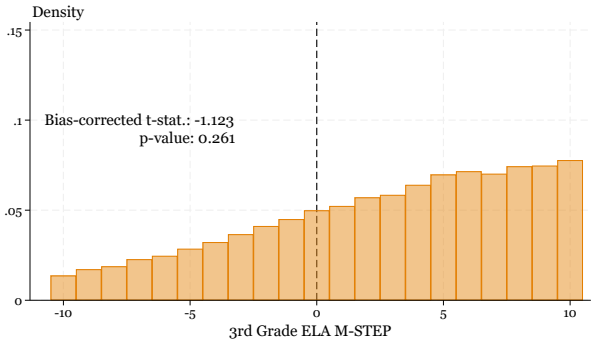
A.2 Additional results

Figure A1: Smoothness of test score distribution density around cut-off

(a) High SES, density of running variable



(b) Low SES, density of running variable



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-centred by the cut-off of 1252. The sample includes students who took the 3rd grade ELA M-STEP for the first time between 2020 and 2022 and scored within 10 points of the cut-off.

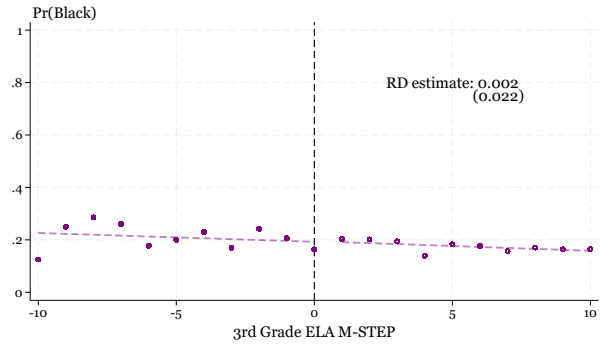
Table A1: M-STEP proficiency levels

	Grade	Level 1 Not proficient	Level 2 Partially proficient	Level 3 Proficient	Level 4 Advanced
ELA	3	1203-1279	1280-1299	1300-1316	1317-1357
Math	3	1217-1280	1281-1299	1300-1320	1321-1361
ELA	4	1301-1382	1383-1399	1400-1416	1417-1454
Math	4	1310-1375	1376-1399	1400-1419	1420-1455

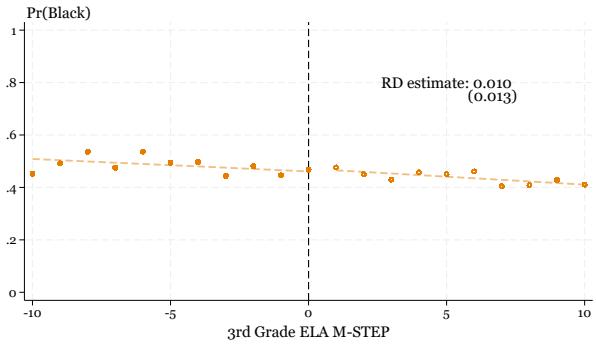
Note: From page 15 of Spring 2023 Interpretive Guide to M-STEP reports (Michigan Department of Education, 2023).

Figure A2: Smoothness of demographics around test score cut-off

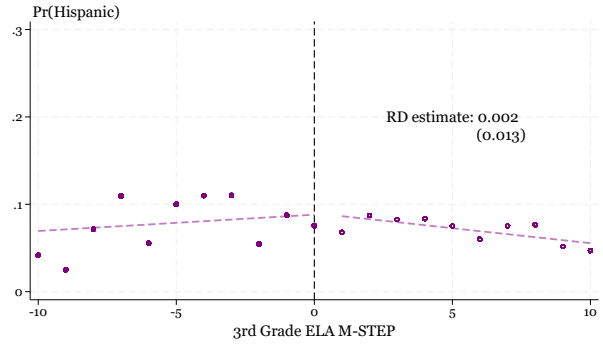
(a) High SES, Black



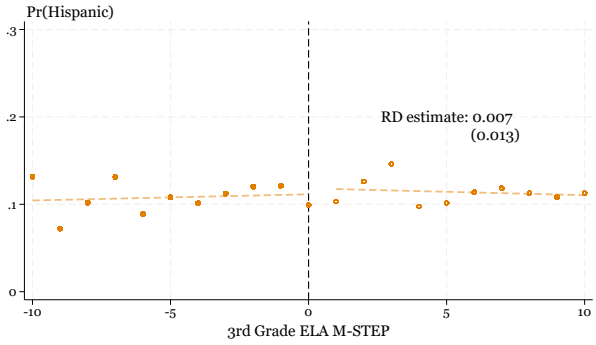
(b) Low SES, Black



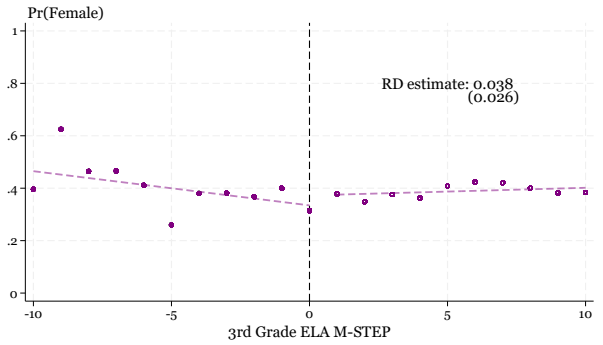
(c) High SES, Hispanic



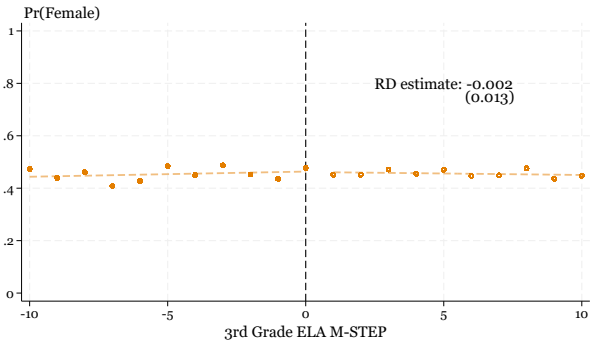
(d) Low SES, Hispanic



(e) High SES, Female



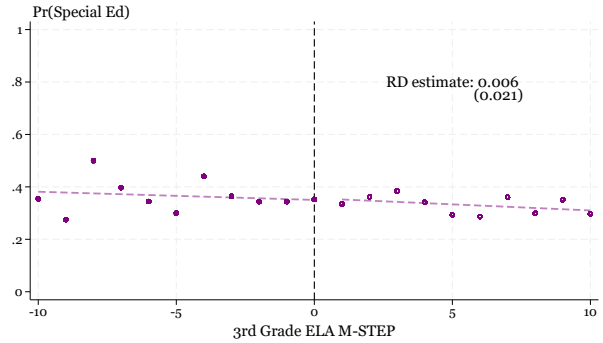
(f) Low SES, Female



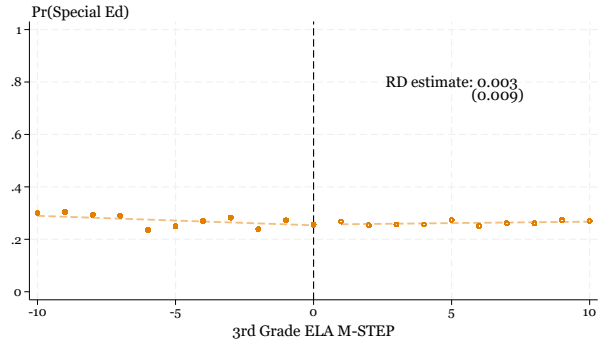
Note: Each Panel presents RD estimates investigating the impact of the RBG3 test score-based promotion policy on the listed baseline covariate, using a local linear specification. The x-axis represents the running variable, the 3rd grade ELA M-STEP, re-centred by the cut-off of 1252. The sample includes students who took the 3rd grade ELA M-STEP for the first time between 2020 and 2022 and scored within 10 points of the cut-off. '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.

Figure A3: Smoothness of additional academic programming around test score cut-off

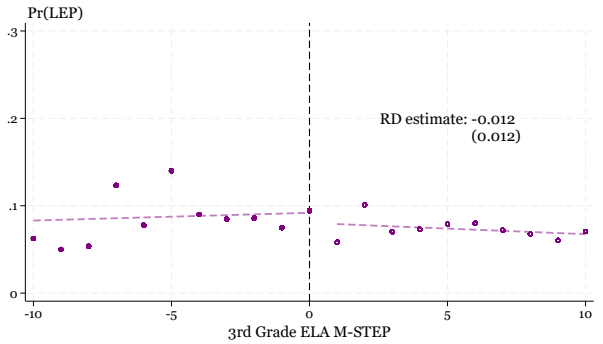
(a) High SES, Special Education



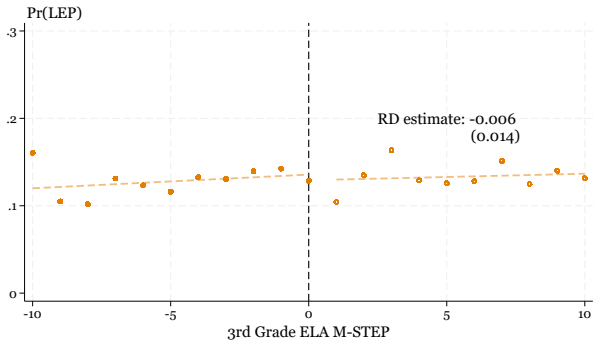
(b) Low SES, Special Education



(c) High SES, Limited English Proficiency



(d) Low SES, Limited English Proficiency



Note: Each Panel presents RD estimates investigating the impact of the RBG3 test score-based promotion policy on the listed baseline covariate, using a local linear specification. The x-axis represents the running variable, the 3rd grade ELA M-STEP, re-centred by the cut-off of 1252. The sample includes students who took the 3rd grade ELA M-STEP for the first time between 2020 and 2022 and scored within 10 points of the cut-off. 'RD estimate' presents β from 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.

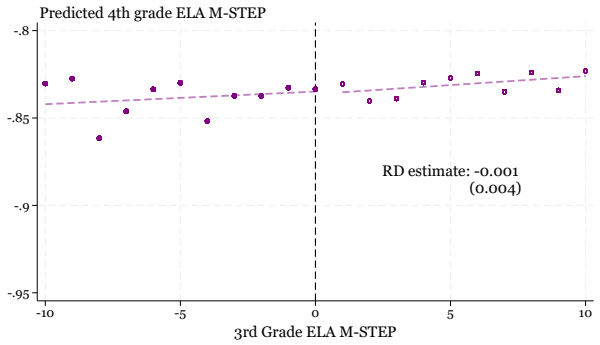
Table A2: Testing effect of test score on promotion and promoted outcomes

	High SES		Low SES	
	Promoted	Proficient	Promoted	Proficient
Above cut-off	0.0194***	-0.0459*	0.0375***	-0.0186**
	(0.0053)	(0.0233)	(0.0043)	(0.0067)
Score	-0.0005	0.0068**	0.0011	0.0034***
	(0.0014)	(0.0025)	(0.0011)	(0.0010)
Above cut-off × Score	0.0011	0.0036	0.0005	0.0014
	(0.0015)	(0.0034)	(0.0012)	(0.0012)
Mean Outcome	0.9703***	0.1483***	0.9330***	0.0653***
	(0.0042)	(0.0142)	(0.0034)	(0.0061)
N	3994.000	3945.000	17796.000	17141.000

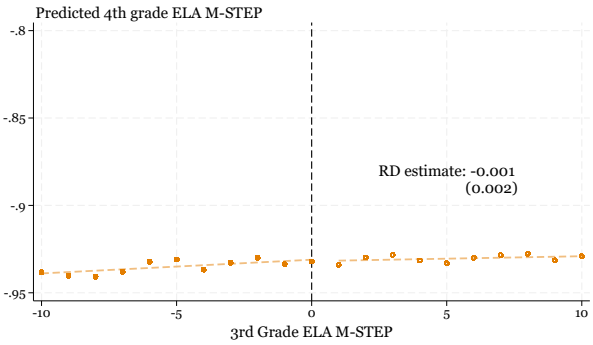
Note: This presents estimates of $X_i = \alpha + \beta\,\mathbb{I}(\text{Score}_i{>}1252) + \delta_1\text{Score}_i + \delta_2\,\mathbb{I}(\text{Score}_i{>}1252) \times \text{Score}_i + \varepsilon_i$, where $X_i$is either an indicator for promotion or the proficiency outcome if promoted. The 'Proficient' columns only include promoted students. Standard errors are clustered at the level of the running variable.

Figure A4: Smoothness of predicted ELA M-STEP next year around test score cut-off

(a) High SES, predicted ELA M-STEP next year



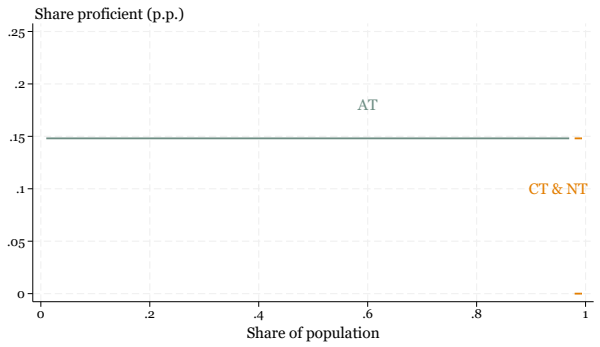
(b) Low SES, predicted ELA M-STEP next year



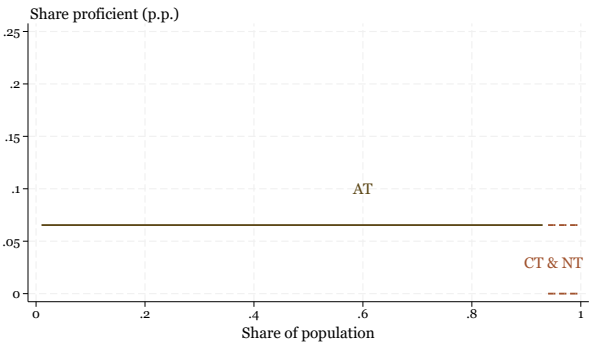
Note: These figures present RD estimates investigating the impact of the RBG3 test score-based promotion policy on the predicted ELA M-STEP score (in standard deviation units) taken in the following school year (regardless of actual promotion status), using demographics, Limited English Proficiency and 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-centred by the cut-off of 1252. The sample includes students who took the 3rd grade ELA M-STEP for the first time between 2020 and 2022 and scored within 10 points of the cut-off.

Figure A5: Average outcomes if promoted by compliance group ($Y_i(1)$)

(a) High SES

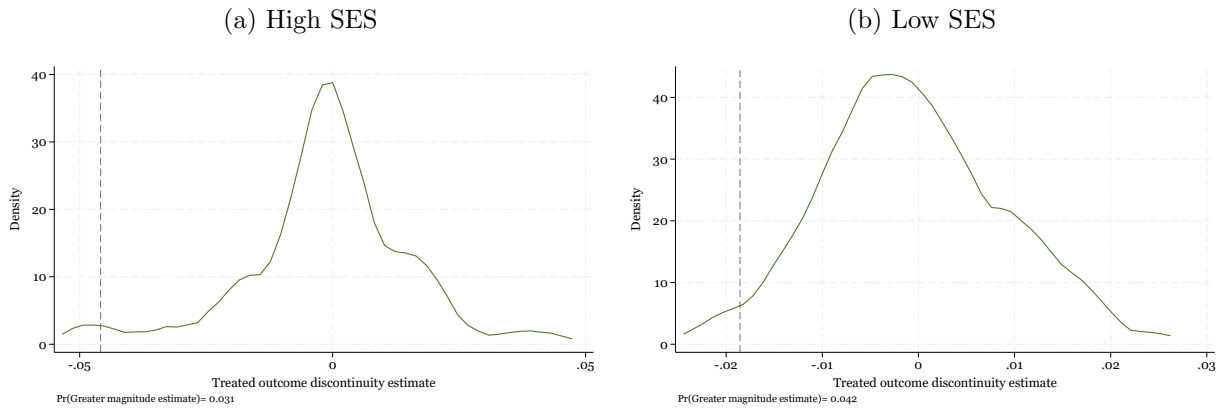


(b) Low SES



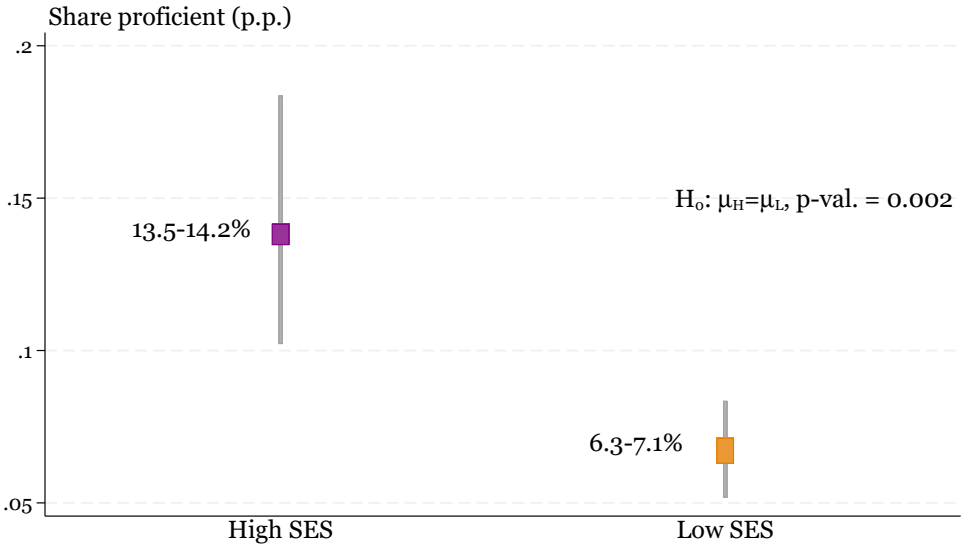
Note: This figure shows the average treated outcomes for always takers ('A'), compliers ('C') and never takers ('N'). The treatment is promotion and the treated outcome, $Y_i(1)$, is whether a student demonstrated any proficiency on both the Math and ELA M-STEP in 4th grade, as defined in Table A1. The 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 A6: Distribution of placebo estimates of discontinuity in treated outcomes



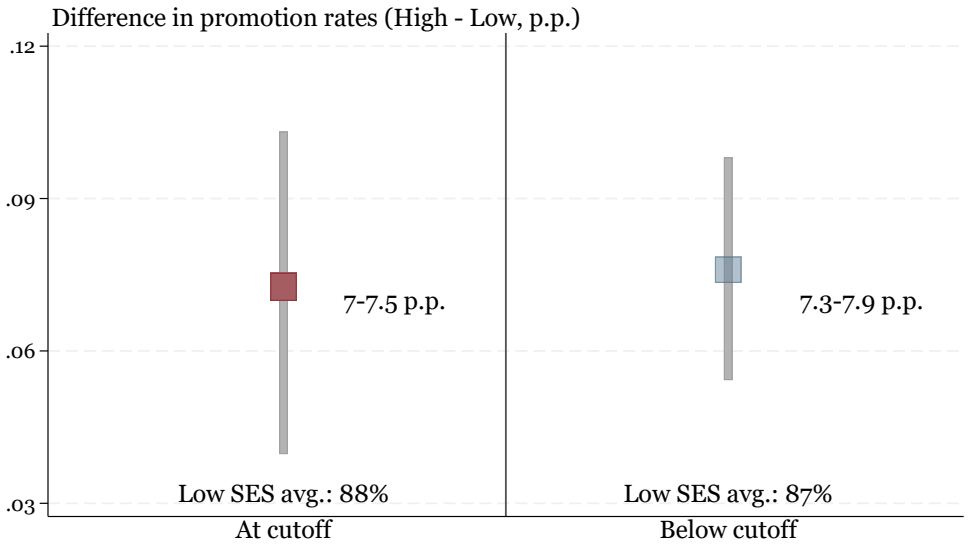
Note: These figures present the distribution of estimates of the discontinuity in treated outcomes obtained by using alternate values of the running variable as the cut-off. We use each value in the [-5,90] interval as the cut-off, where 0 is the true cut-off, but keep the bandwidth fixed at 10 points on either side of the cut-off, as in the baseline. The vertical lines display the observed RD estimate that we obtain in Figure 3. 'Pr(Greater magnitude estimate)' computes the share of placebo estimates that are larger than the observed RD estimate. 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 in Table A1. The sample and specification details are the same as described in Figure 3.

Figure A7: Average outcomes if promoted, excluding students whose parents requested exemptions



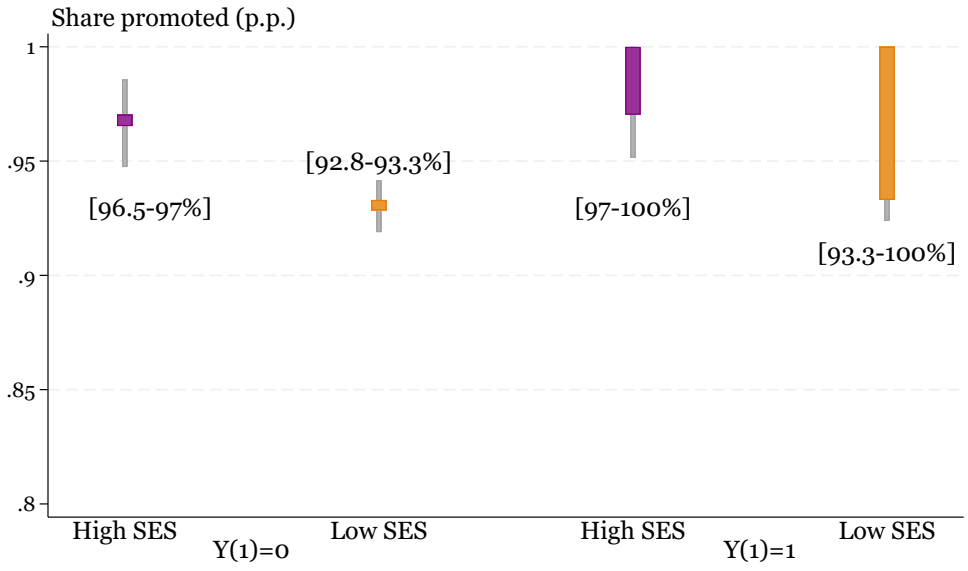
Note: This figure presents bounds on the average treated outcome obtained using the approach described in Section 3. The treatment is promotion and the treated outcome, $Y_i(1)$, is whether a student demonstrated any proficiency on both the Math and ELA M-STEP in 4th grade, as defined in Table A1. Confidence 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 C.5.

Figure A8: SES promotion gap, excluding students whose parents requested exemptions



Note: This figure presents bounds on the average difference in promotion rates, conditional on treated potential outcomes, using the approach described in Section 3. 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 in Table A1. Confidence intervals are for the true parameter and are bootstrapped using 1,000 replications and a Bayesian bootstrap.

Figure A9: Promotion rates at the cut-off, conditional on promoted outcome (π_{0ry})



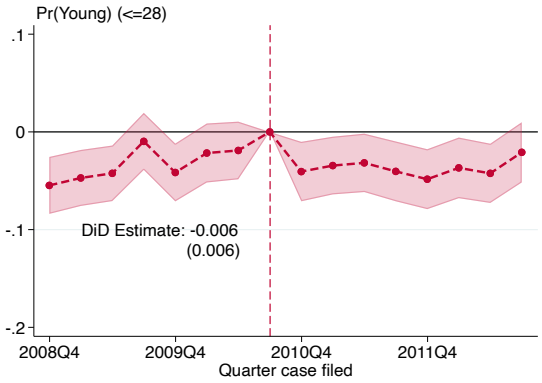
Note: This figure presents bounds on the average prosecution rates for each SES group, conditional on promoted outcomes, $Y_i(1)$, using the approach described in Section 3. $Y_i(1)$ is whether a student demonstrated any proficiency on both the Math and ELA M-STEP in 4th grade, as defined in Table A1. Confidence intervals are for the true parameter and are bootstrapped using 1,000 replications and a Bayesian bootstrap.

Appendix B Additional results: Misdemeanor prosecution

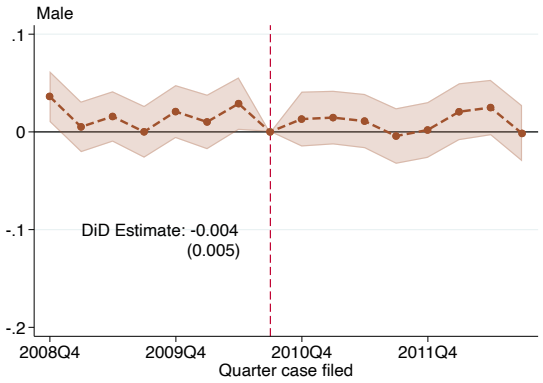
B.1 Robustness checks

Figure B1: Impact of budget reform on caseload characteristics

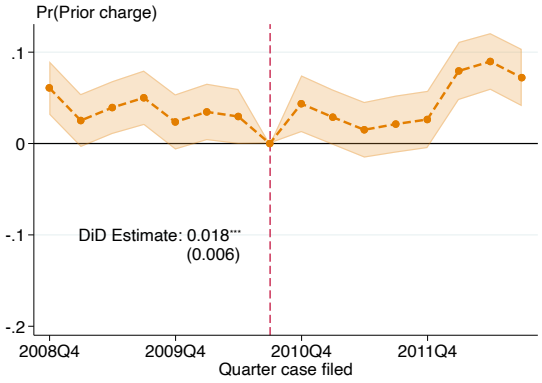
(a) Age $\leq$ 28



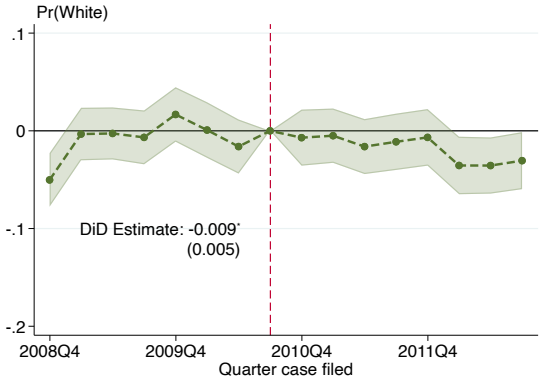
(b) Male



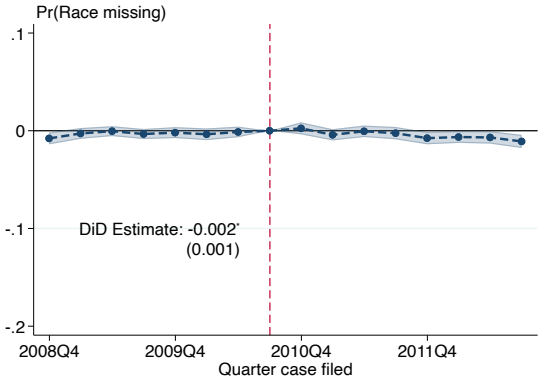
(c) Any prior charge



(d) White (Non-Hispanic)



(e) Missing race/ethnicity

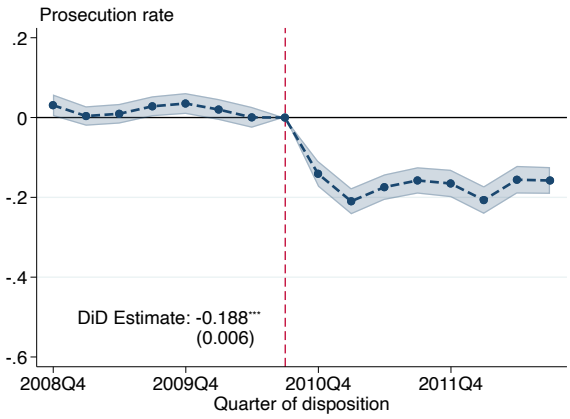


Note: Each Panel presents event study estimates investigating the impact of the King County budget reform. 'DiD estimate' is β^{DD} from $X_{igt} = \alpha + \delta_1 \mathbb{I}[\text{King County}] + \delta_2 Post_i + \beta^{DD} \mathbb{I}[\text{King County}] \times Post_i + \epsilon_{igt}$, where $Post_i = 1$ if the case is filed on or after September 28, 2010, when the budget reform was announced, where X_{igt} denotes a baseline characteristic. 'Young' includes defendants who less than 29 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 B2: Robustness of first stage to more expansive prosecution decision

(a) White defendants

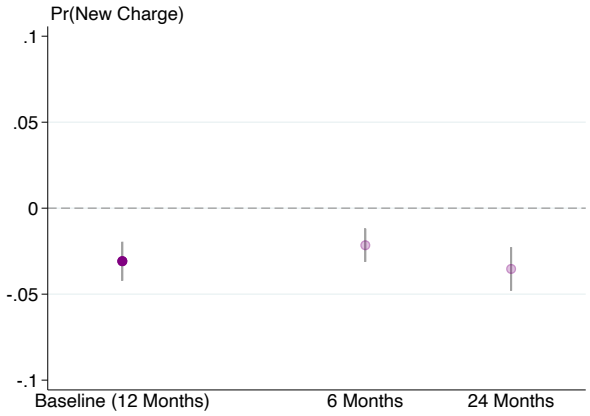
(b) Minority defendants



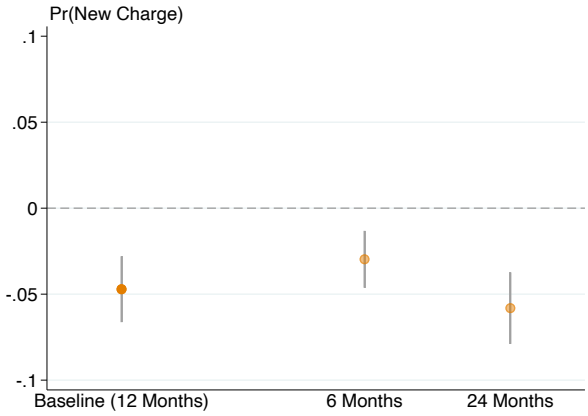
Note: This identical to Figure 6 except includes cases listed as 'Dismissed' but with a fine as part of the sentence as prosecuted.

Figure B3: Impact of King County budget reform on re-offence rates over different time horizons

(a) White defendants



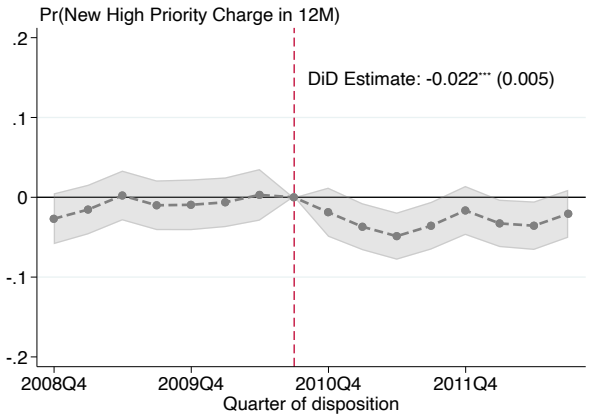
(b) Minority defendants



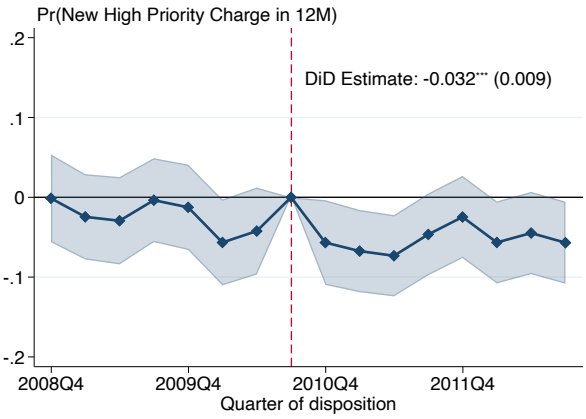
Note: Each Panel presents DiD estimates as reported in Figure 7, but for different amounts of time after disposition.

Figure B4: Impact of King County budget reform on 'high-priority' re-offence within one year

(a) White defendants



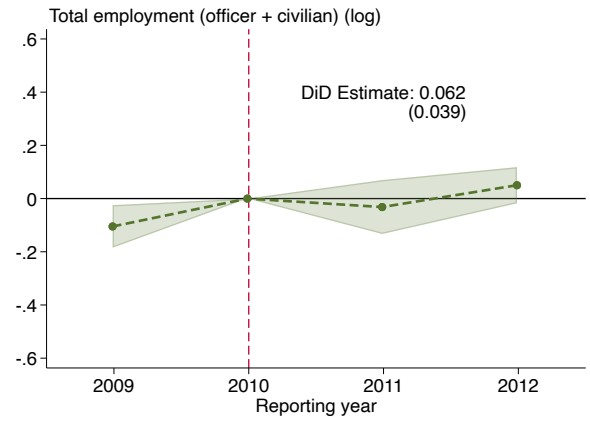
(b) Minority defendants



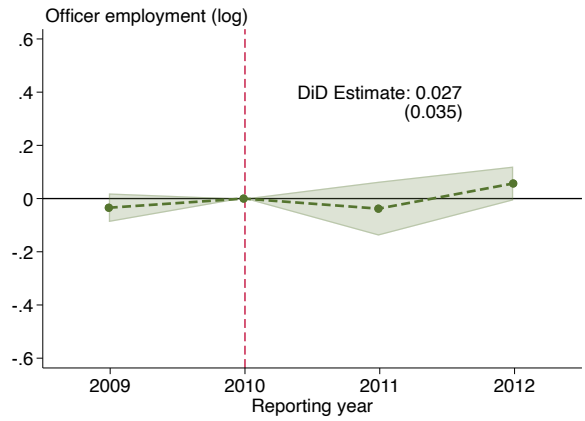
Note: Each Panel presents event study estimates investigating the impact of the King County budget reform. The re-offence outcome includes any 'high-priority' charges filed against an individual anywhere in Washington State. 'High-priority' cases are those that are not associated with charges that were commonly dismissed in the 2 quarters after the budget reform. Sample includes all misdemeanor defendants, as described in Table 2. '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 of on or after September 28, 2010, when the budget reform was announced. 95% confidence intervals are constructed using heteroscedasticity-robust standard errors.

Figure B5: Impact of budget reform on police employment

(a) Total employment



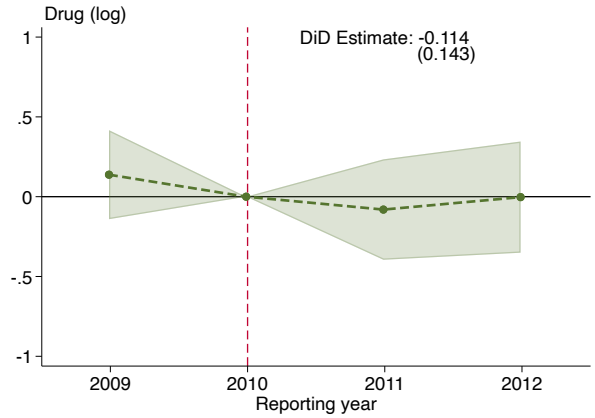
(b) Officer employment



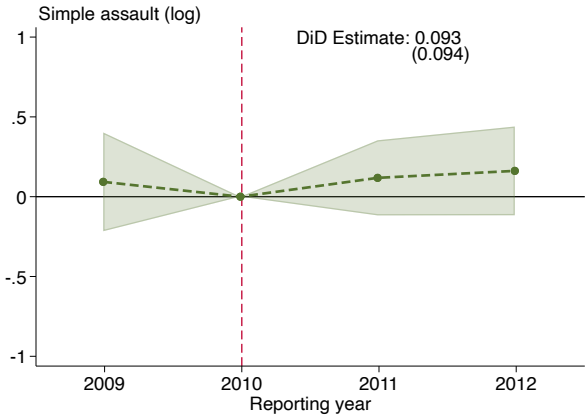
Note: Estimates generated using annual average employment from the Law Enforcement Officers Killed and Assaulted (LEOKA) data (Kaplan, 2023). This specification is similar to those used to estimate the first stage and reduced form except for the inclusion of originating agency (ORI) fixed effects and 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 B6: Impact of budget reform on arrests

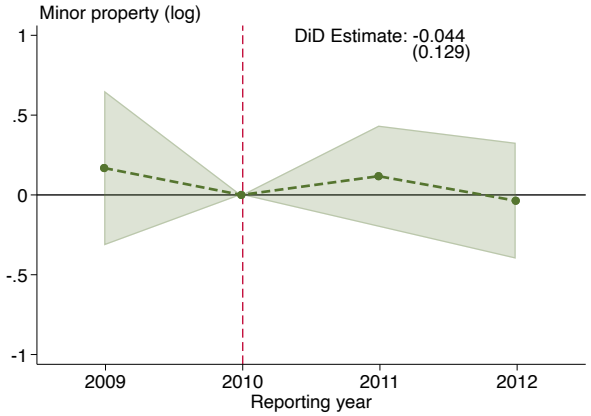
(a) Drug (possession +sale)



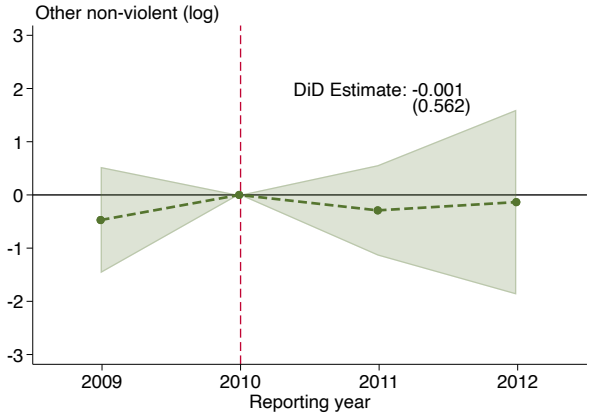
(b) Simple assault



(c) Minor property



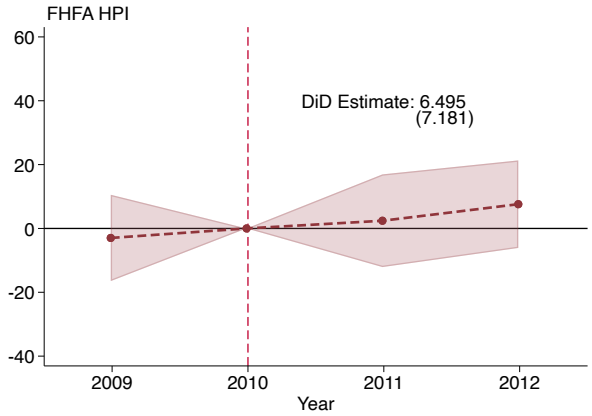
(d) Minor other arrests



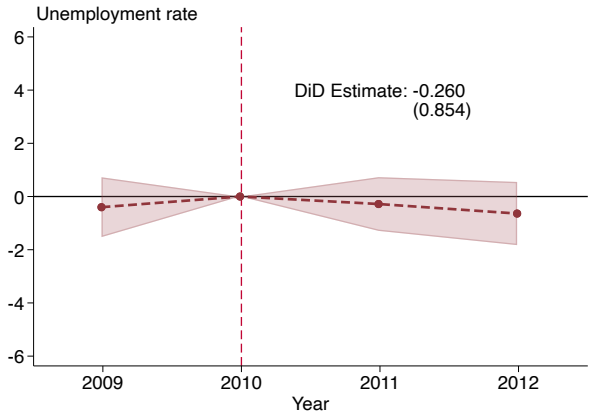
Note: Estimates generated using annual number of arrests from Uniform Crime Report (UCR) data (Kaplan, 2023). 'Minor property' arrests include arrests for stolen property, fraud, forgery and theft. This specification is similar to those used to estimate the first stage and reduced form except for the 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 B7: Impact of budget reform on non-crime economic factors

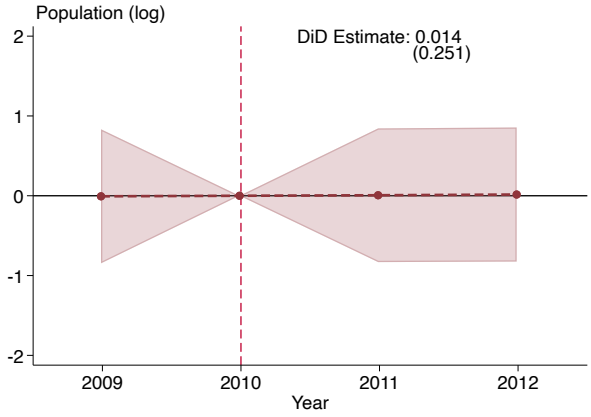
(a) FHFA House Price Index



(b) Unemployment rate



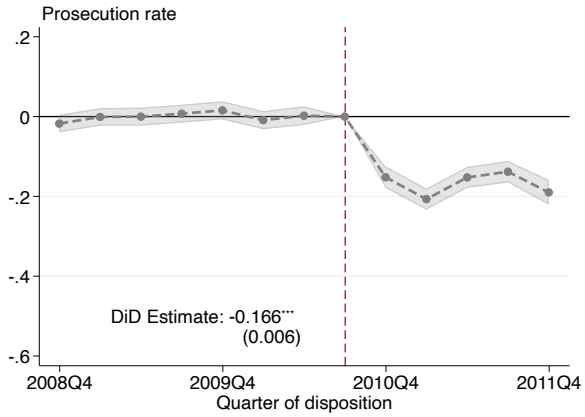
(c) Population (log)



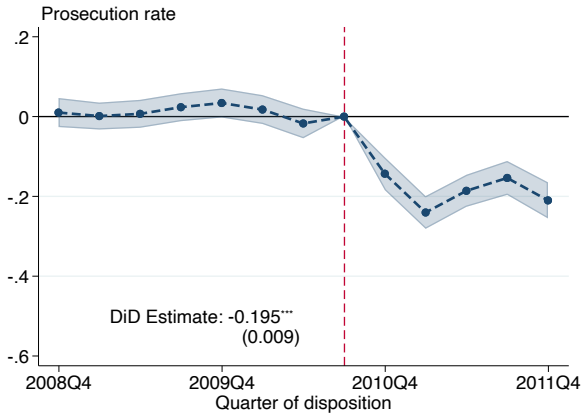
Note: This specification used here is similar to those used to estimate the first stage and reduced form except for the usage of the average pre-reform county-level population as weights in Panels (a) and (b). County-level data on house prices are from the Federal Housing Finance Agency (FHFA), unemployment 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 B8: Robustness of first stage and reduced form to excluding post-SPD investigation period

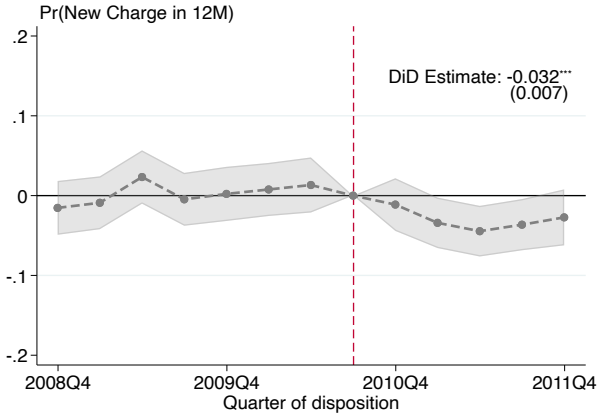
(a) First stage, White defendants



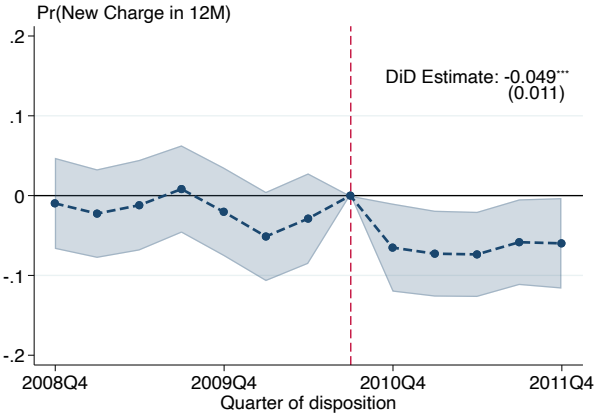
(b) First stage, Minority defendants



(c) Reduced form, White defendants



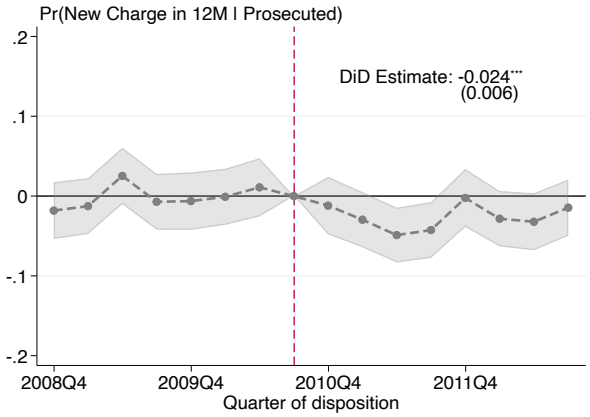
(d) Reduced form, Minority defendants



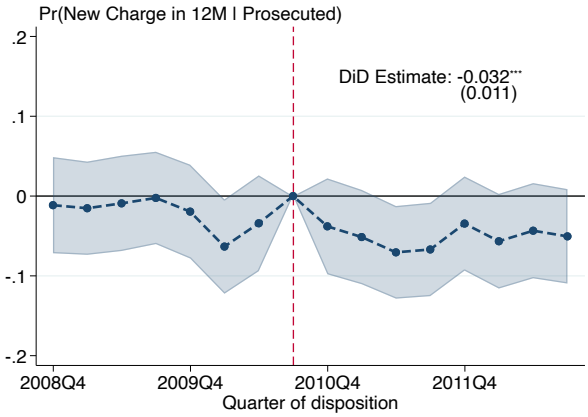
Note: This identical to Figure 6 and Figure 7 except excludes cases disposed of after December 11, 2011.

Figure B9: Impact of King County budget reform on re-offence within one year, only prosecuted defendants

(a) White defendants

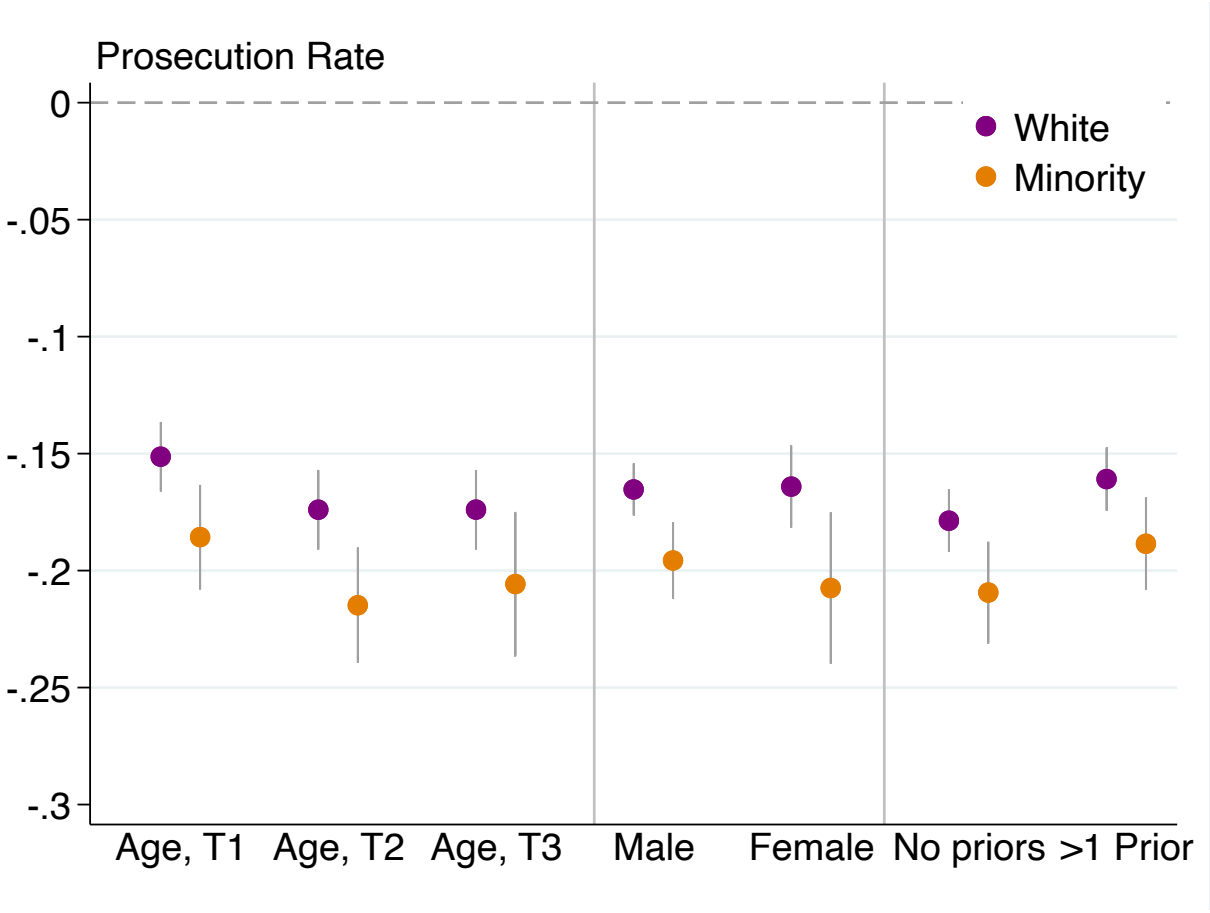


(b) Minority defendants



Note: Each Panel is identical to Figure 7, except the sample only includes prosecuted defendants.

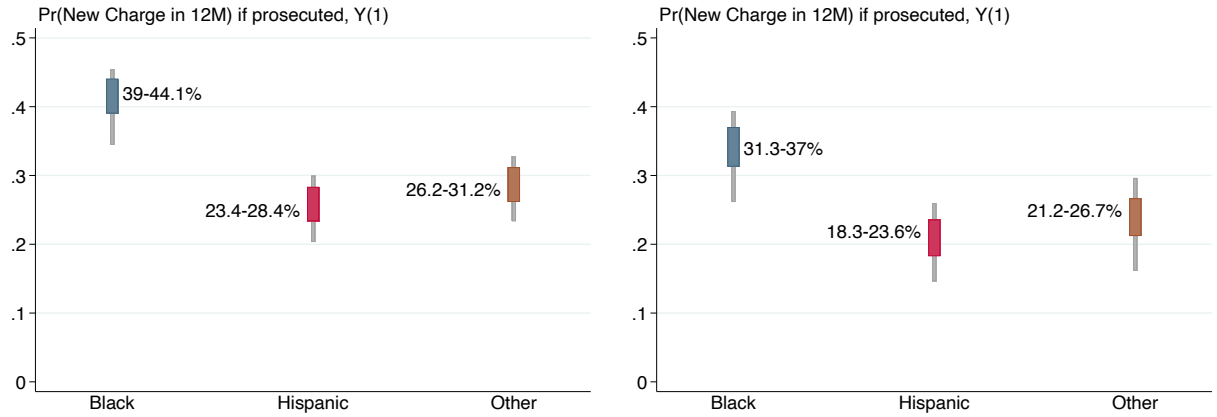
Figure B10: Testing for defiers: First stage by subgroup



Note: Each Panel presents DiD estimates as reported in Figure 6, but for different covariate subgroups. Age at disposition is split into terciles, represented by T1–T3.

Figure B11: Average outcomes if prosecuted, disaggregated by minority subgroup

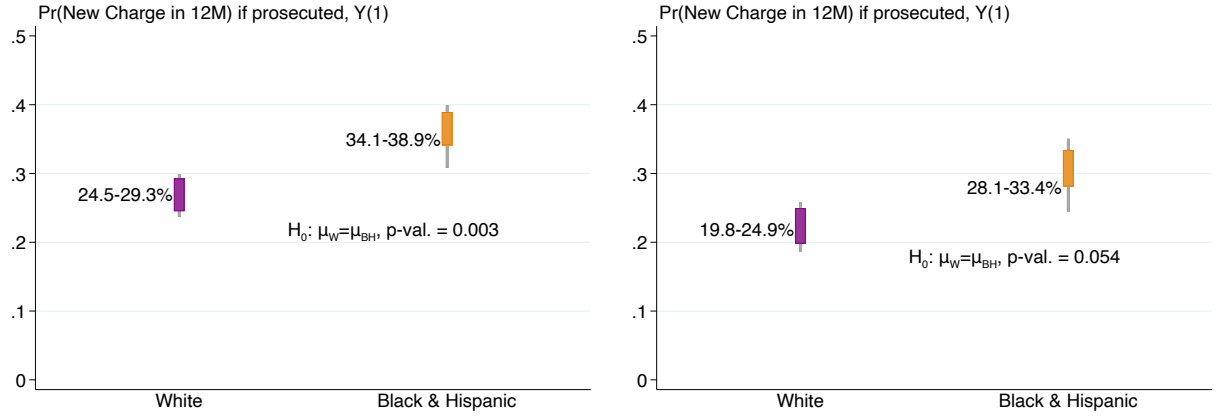
(a) Pre-reform ($Z = 0$) (b) Post-reform ($Z = 1$)



Note: This figure presents bounds on the average treated outcome obtained using the approach described in Section 3, separately by time period and subgroups within minority defendants. The treatment is prosecution and the treated outcome, $Y_i(1)$, is whether an individual is charged with a new offence within one year after disposition, if prosecuted. Confidence intervals are for the true parameter and are bootstrapped using 1,000 replications and a Bayesian bootstrap.

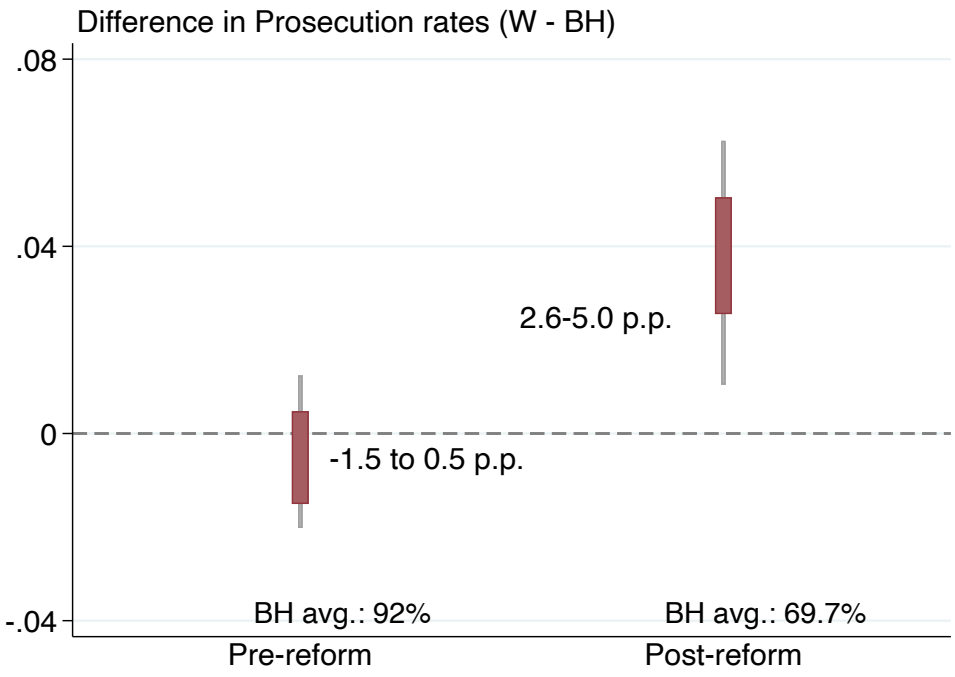
Figure B12: Average outcomes if prosecuted: White vs. Black/Hispanic defendants

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



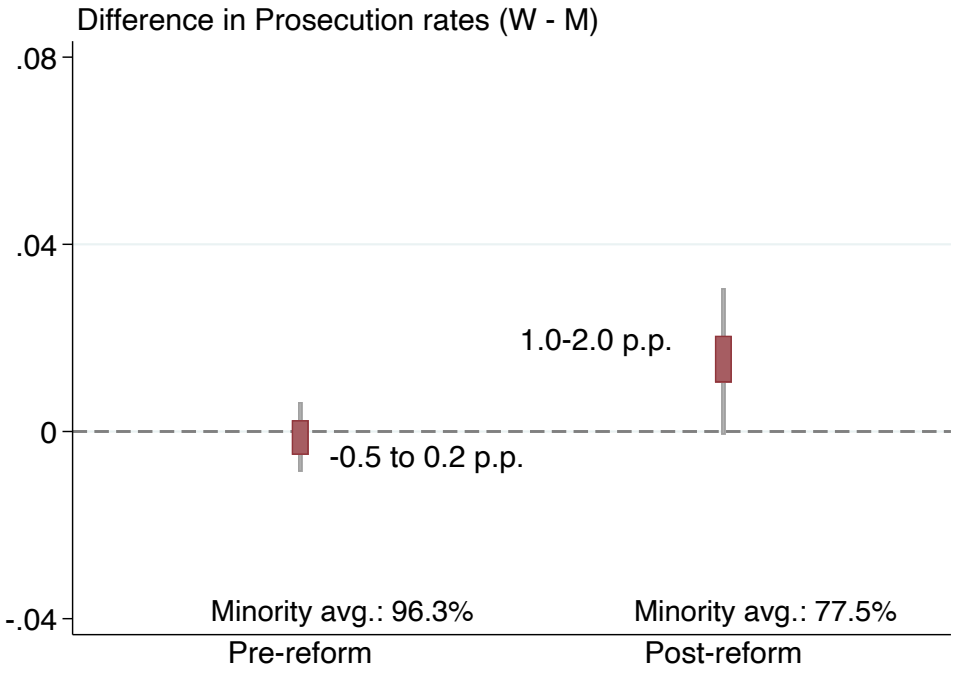
Note: This figure presents bounds on the average treated outcome obtained using the approach described in Section 3, separately by race and time period. The treatment is prosecution and the treated outcome, $Y_i(1)$, is whether an individual is charged with a new offence within one year after disposition, 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 C.5.

Figure B13: 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 3. $Y_i(1)$ is whether an individual is charged with a new offence within one year after disposition, if prosecuted. Confidence intervals are for the true parameter and are bootstrapped using 1,000 replications and a Bayesian bootstrap.

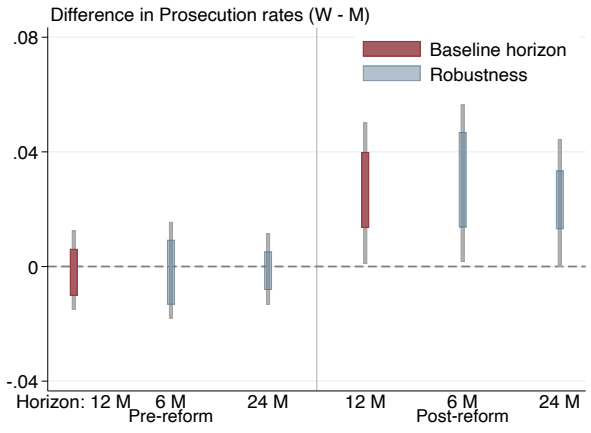
Figure B14: Racial prosecution gap cond. on prosecuted outcome: Broad prosecution definition



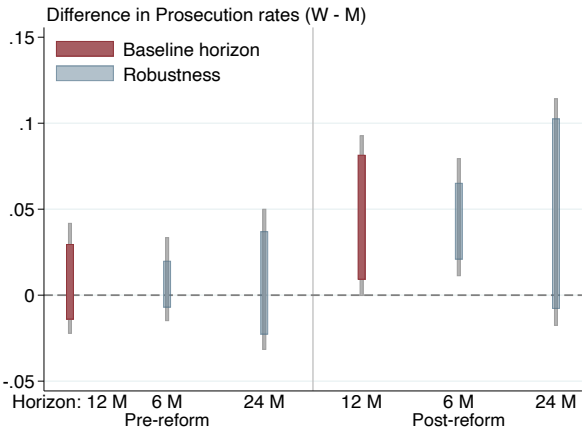
Note: This figure presents bounds on the average difference in prosecution rates in each time period, conditional on prosecuted outcomes, $Y_i(1)$, using the approach described in Section 3. $Y_i(1)$ is whether an individual is charged with a new offence within one year after disposition, if prosecuted. Confidence intervals are for the true parameter and are bootstrapped using 1,000 replications and a Bayesian bootstrap.

Figure B15: Racial prosecution gap conditional on potential outcomes: Variation by outcome horizon

(a) $Y_i(1)$

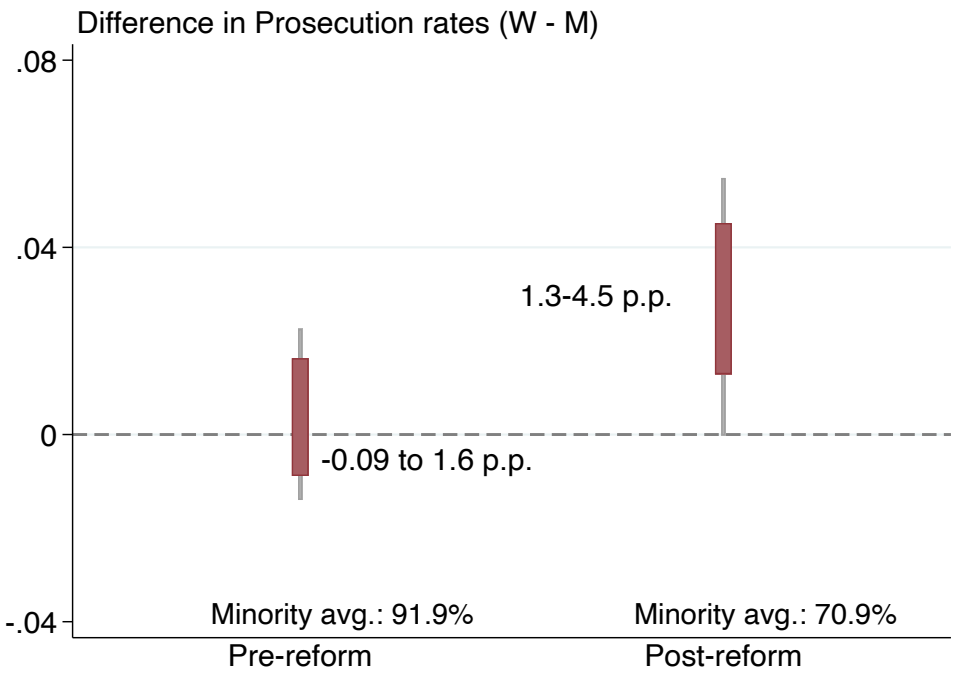


(b) $Y_i(0)$



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 is charged with a new offence 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.

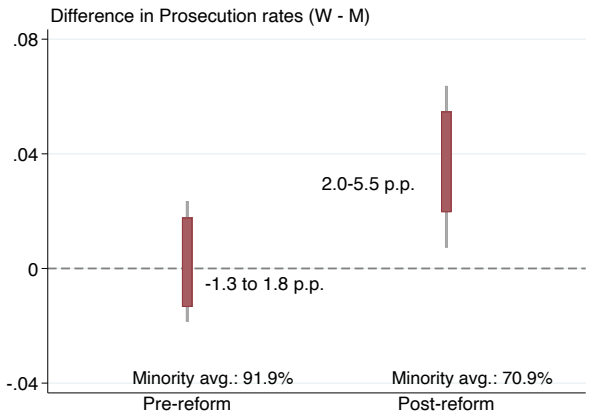
Figure B16: Racial prosecution gap conditional on prosecuted outcome: Worst case bounds



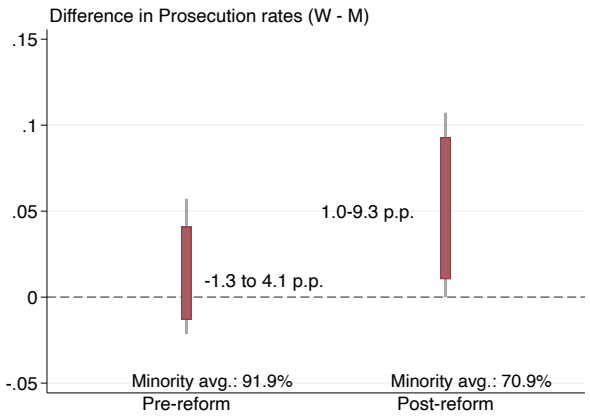
Note: This figure presents bounds on the average difference in prosecution rates in each time period, conditional on prosecuted outcomes, $Y_i(1)$, assuming that never takers' outcomes are bounded between 0 and 1. $Y_i(1)$ is whether an individual is charged with a new offence within one year after disposition,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

Figure B17: Racial prosecution gap, conditional on potential outcomes and baseline covariates

(a) $Y_i(1)$

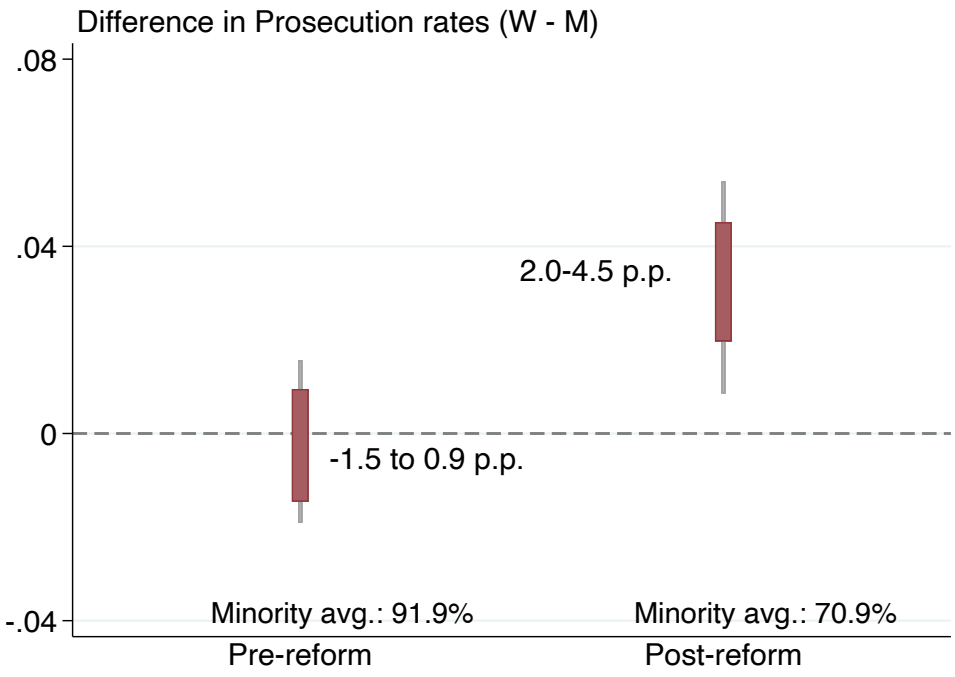


(b) $Y_i(0)$



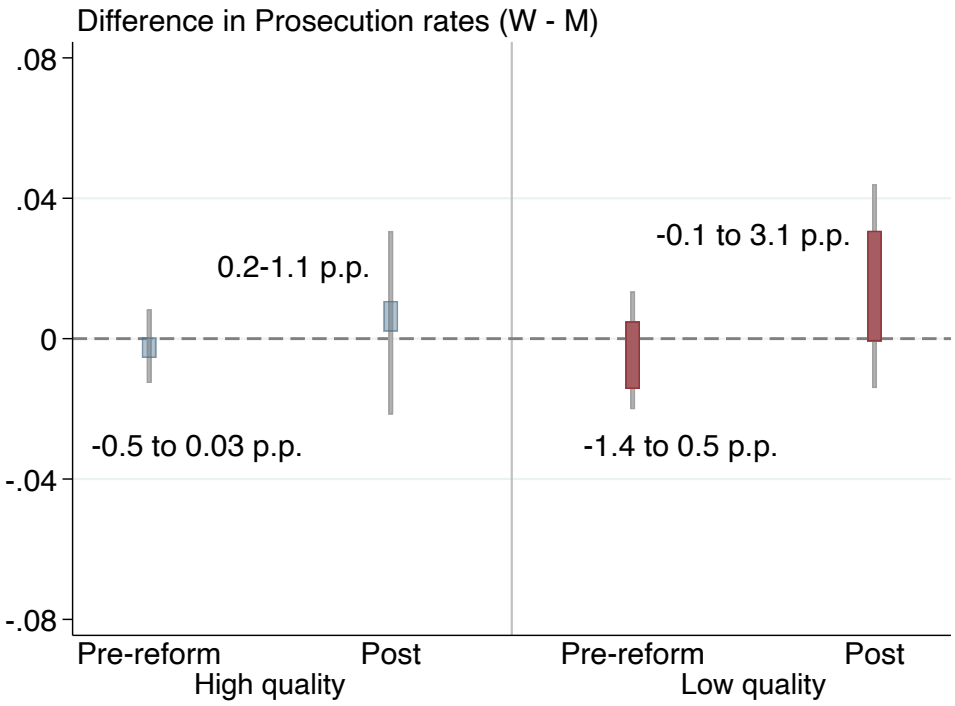
Note: This figure presents bounds on the average difference in prosecution rates in each time period, conditional on potential outcomes, $Y_i(\cdot)$, and baseline covariates. $Y_i(\cdot)$ is whether an individual is charged with a new offence within one year after disposition if prosecuted (Panel (a)) or if dismissed (Panel (b)). Confidence intervals are for the true parameter and are bootstrapped using 1,000 replications and a Bayesian bootstrap.

Figure B18: Racial prosecution gap conditional on receiving any punishment if prosecuted



Note: This figure presents bounds on the average difference in prosecution rates in each time period, conditional on case outcome if prosecuted, $Y_i(1)$, using the approach described in Section 3. $Y_i(1)$ is whether an individual is sentenced to any non-fine punishment, if prosecuted. Confidence intervals are for the true parameter and are bootstrapped using 1,000 replications and a Bayesian bootstrap.

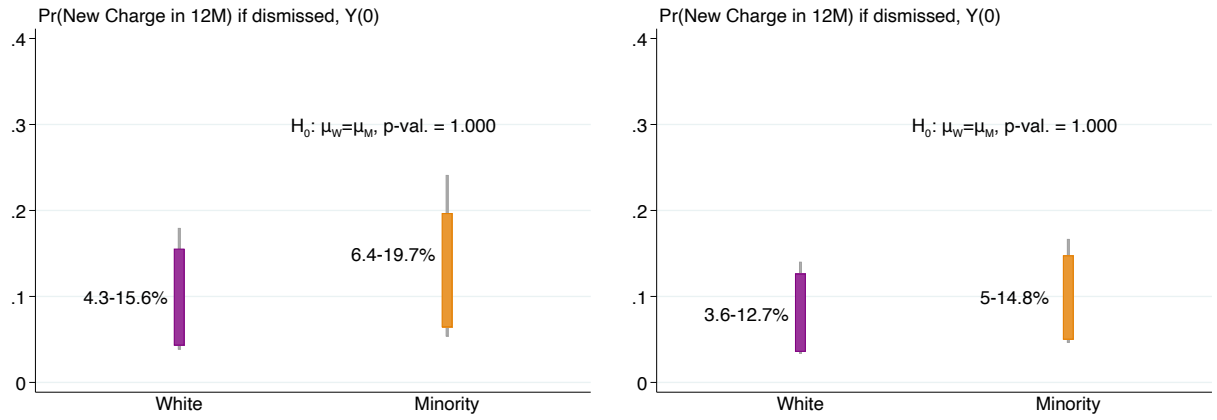
Figure B19: Racial prosecution gap, by proxy for case quality: excluding cases without charges



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 3. $Y_i(1)$ is whether an individual is charged with a new offence within 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.

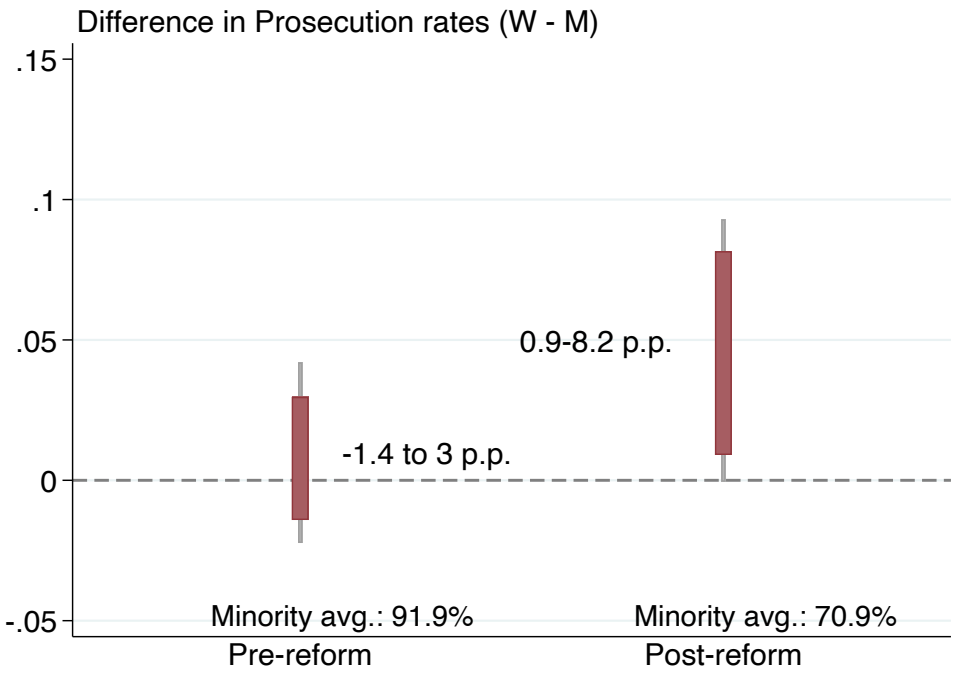
Figure B20: Average outcomes if dismissed, $E[Y_{it}(0)|R_i = r]$

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



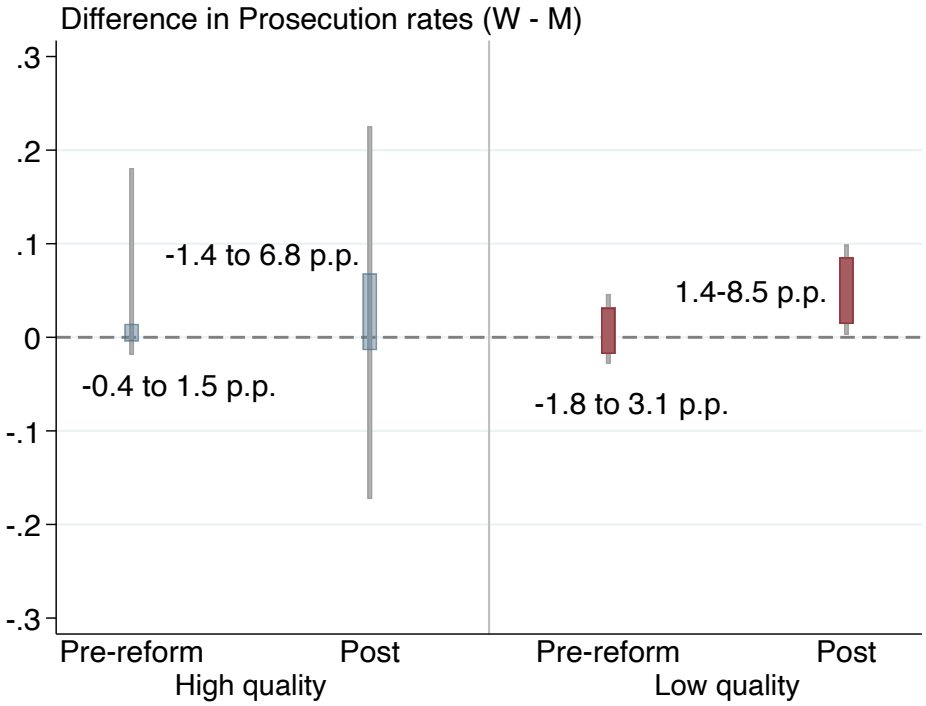
Note: This figure presents bounds on the average untreated outcome obtained using the approach described in Section 3, separately by race and time period. The treatment is prosecution and the untreated outcome, $Y_i(0)$, is whether an individual is charged with a new offence within one year after disposition, if dismissed. Confidence intervals are for the true parameter and are bootstrapped using 1,000 replications and a Bayesian bootstrap.

Figure B21: 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 3. $Y_i(0)$ is whether an individual is charged with a new offence within 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 B22: Racial prosecution gap, by proxy for case quality: conditional on dismissed outcome

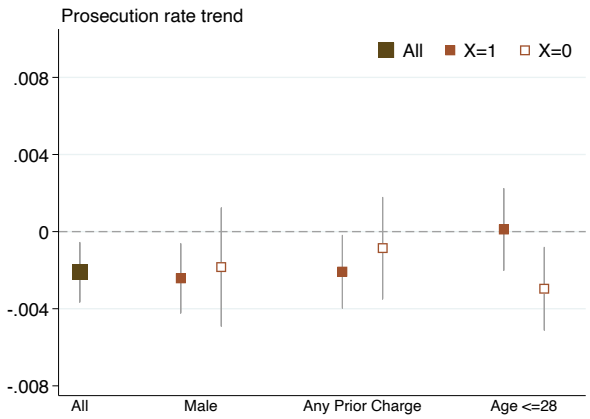


Note: This figure presents bounds on the average difference in prosecution rates in each time period, conditional on dismissed outcomes, $Y_i(0)$, using the approach described in Section 3. $Y_i(0)$ is whether an individual is charged with a new offence within one year after disposition, if dismissed. '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.

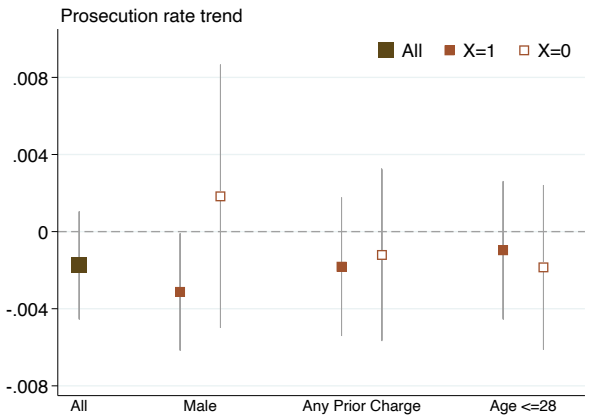
B.2 Empirically validating DiD adjustment assumptions

Figure B23: 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 in the counties adjacent to King County, among the subgroup, where X denotes whether the dummy is equal to 1 or not. The estimate labelled 'All' reproduces the overall trend estimate from Figure 9. 95% confidence intervals are constructed using heteroscedasticity-robust standard errors.

Figure B24 and Figure B26 estimate the following regression & tests the hypotheses listed below it:

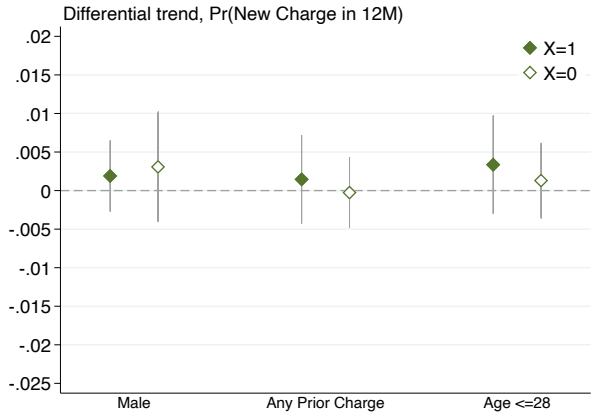
$$
\begin{aligned}
Y_{itg} = {} & \beta_1 \mathrm{t} + \beta_2 X_i + \beta_3 \text{King County} + \\
& \delta_1 X_i \times \mathrm{t} + \delta_2 X_i \times \text{King County} + \delta_3 \mathrm{t} \times \text{King County} + \\
& \delta_4 X_i \times \mathrm{t} \times \text{King County} + \varepsilon_{igt} \\
& X = 0 : H_0 : \delta_3 = 0 \\
& X = 1 : H_0 : \delta_3 + \delta_4 = 0
\end{aligned}
$$

Figure B25 and Figure B27 estimates the following regression & tests the hypotheses listed below it:

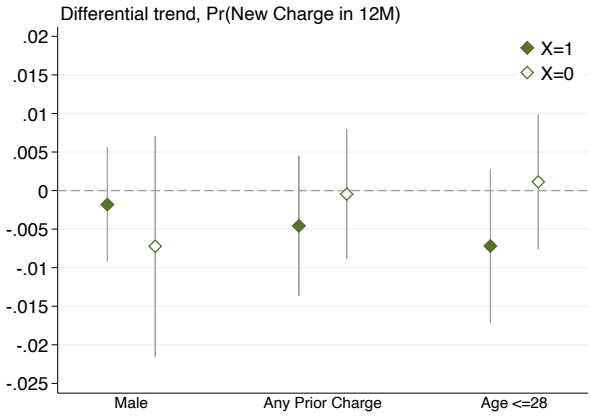
$$
\begin{aligned}
Y_{itg} = {} & \beta_1 \mathrm{t} + \beta_2 X_i + \beta_3 \text{King County} + \\
& \delta_1 X_i \times \mathrm{t} + \delta_2 X_i \times \text{King County} + \delta_3 \mathrm{t} \times \text{King County} + \\
& \delta_4 X_i \times \mathrm{t} \times \text{King County} + \varepsilon_{igt} \\
& \text{King County } H_0 : \delta_1 + \delta_4 = 0 \\
& \text{Adjacent } H_0 : \delta_1 = 0
\end{aligned}
$$

Figure B24: Testing for differential trends in group-specific treated outcomes across counties

(a) White defendants



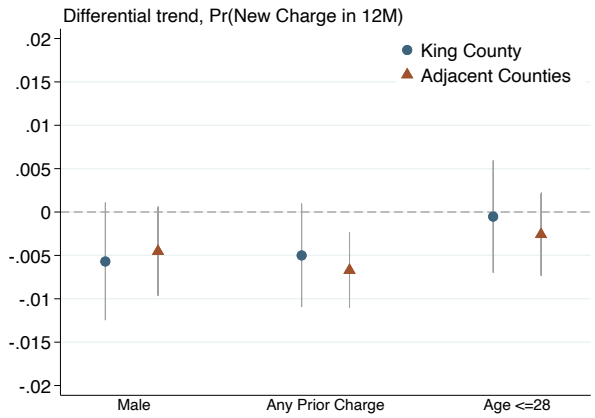
(b) Minority defendants



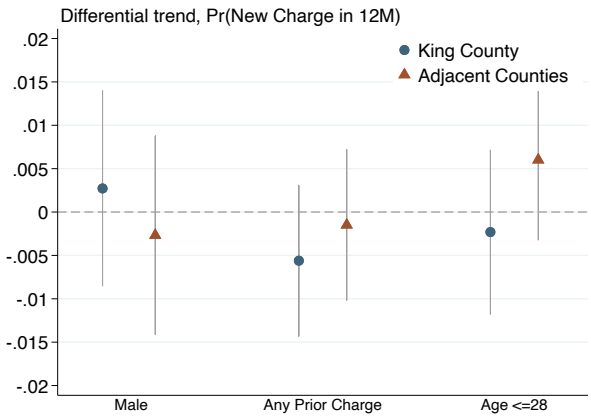
Note: Each coefficient is an estimate of the difference in pre-period trends in treated outcomes (outcomes if prosecuted) across counties, for a given covariate value. The sample only includes individuals who are prosecuted prior to the budget reform. For example, the first green diamond in Panel a) 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 B25: Testing for differential trends in treated outcomes between groups, within each county

(a) White defendants



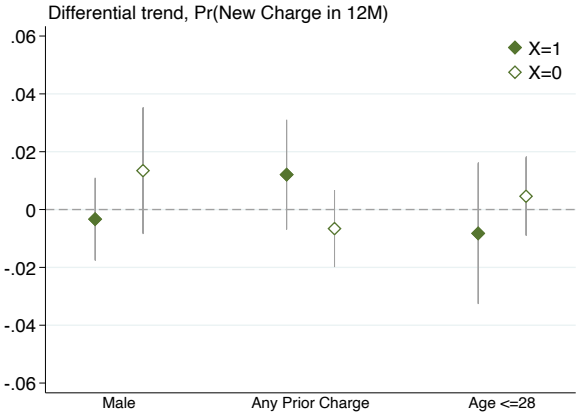
(b) Minority defendants



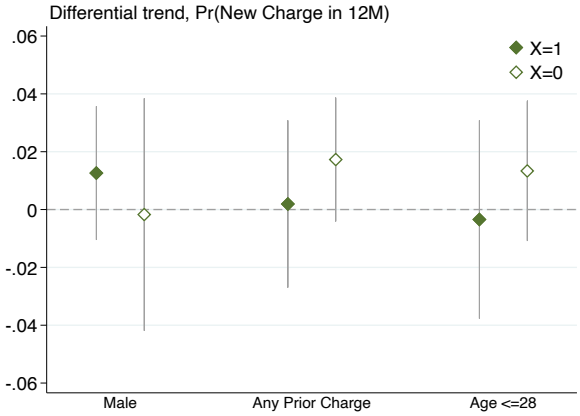
Note: Each coefficient is an estimate of the difference in pre-period trends in treated outcomes (outcomes if prosecuted) across covariate groups, within a given county. The sample only includes individuals who are prosecuted prior to the budget reform. For example, the first blue circle in Panel a) 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 B26: Testing for differential trends in group-specific untreated outcomes across counties

(a) White defendants



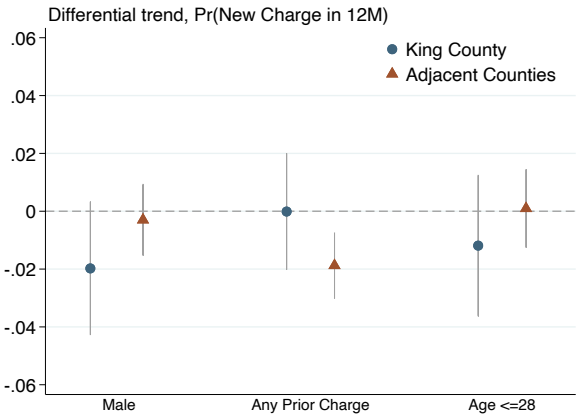
(b) Minority defendants



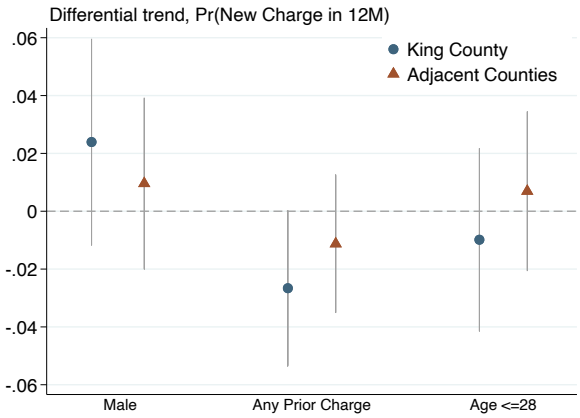
Note: Each coefficient is an estimate of the difference in pre-period trends in untreated outcomes (outcomes if not prosecuted) across counties, for a given covariate value. The sample only includes individuals who are dismissed prior to the budget reform. For example, the first green diamond in Panel 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 B27: Testing for differential trends in untreated outcomes between groups, within each county

(a) White defendants



(b) Minority defendants



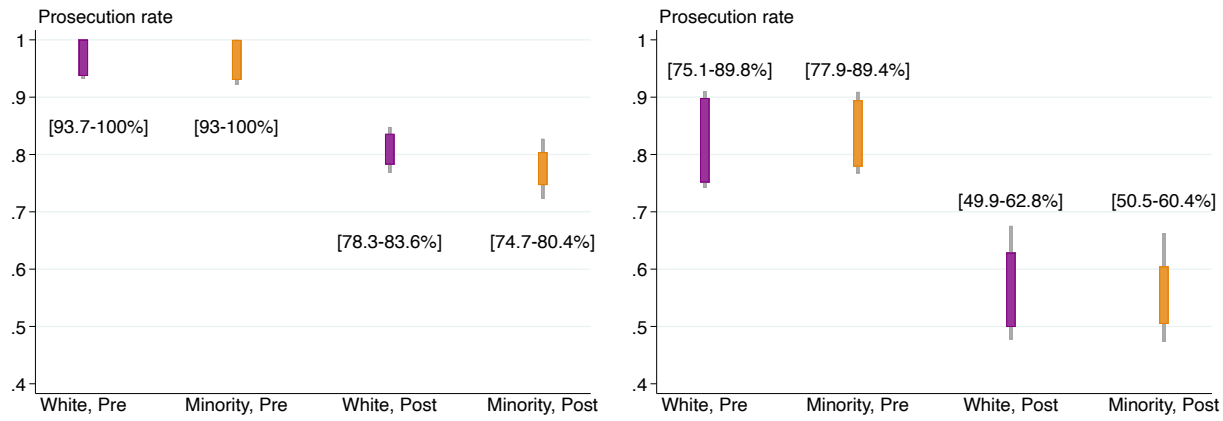
Note: Each coefficient is an estimate of the difference in pre-period trends in untreated outcomes (outcomes if not prosecuted) across covariate groups, within a given County. The sample only includes individuals who are dismissed prior to the budget reform. For example, the first blue circle in Panel 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.

B.3 How prosecution varies by potential outcomes & treatment effects

Here we discuss how the race-, time period- and potential outcome-specific prosecution rates, which are the building blocks of our discrimination estimates (see Equation 3), provide some evidence on prosecution decisions in this setting. We start by conditioning on re-offence outcomes if prosecuted, and Figure B28 presents estimates of prosecution rates separately by race, time period, and whether individuals **would not** commit a new offence if prosecuted ($Y_i(1) = 0$) or **would** commit a new offence if prosecuted ($Y_i(1) = 1$).[60]

Figure B28: Prosecution rates conditional on prosecuted outcome (π_{zry})

(a) No new offence if prosecuted, $Y_{it}(1) = 0$ (b) New offence if prosecuted, $Y_{it}(1) = 1$



Note: This figure presents bounds on the average prosecution rates for each race group and time period, conditional on prosecuted outcomes, $Y_i(1)$, using the approach described in Section 3. $Y_i(1)$ is whether an individual is charged with a new offence within one year after disposition, if prosecuted. Confidence intervals are for the true parameter and are bootstrapped using 1,000 replications and a Bayesian bootstrap.

Panel a) of Figure B28 focuses on prosecution rates for individuals who would not commit a new offence if prosecuted ($Y_i(1) = 0$). The first two estimates are prosecution rates before the reform ("Pre") and the following two estimates are prosecution rates after the reform ("Post"). We see that among defendants who would not commit a new offence if prosecuted, prosecution rates before the reform are similar across race. However, we see suggestions that the drop after the reform is greater for minority defendants. Prosecution rates for white defendants falls from 93.7%–100% to between 78.3%–83.6%. Prosecution rates for minority defendants fall from a similar pre-reform level to 74.7%–80.4%. This is not the case in Panel b), which displays the analogous prosecution rates for individuals who would commit a new offence if they were prosecuted. Here, prosecution falls evenly by racial group. These patterns suggest that the post-reform racial gap that we see in Figure 12 is concentrated among individuals who would not commit an offence if prosecuted.

[60] Using the bounds on prosecution rates in Figure B28 to construct average discrimination conditional on prosecuted outcomes following Equation 3 will not yield the same estimates as the main analysis, which plugs the bounds for the average prosecuted outcomes directly into the expression for our main object of interest, the average period-specific racial gap, $\triangle_z$. The two approaches involve computing averages and minima/maxima in different orders, which will produce numerically different results because minimum/maximum are not linear functions. See Section 3 for additional details.

These patterns also provide suggestive evidence that prosecution is targeted based on the potential outcomes of defendants. Specifically, we see that prosecutors are less likely to prosecute individuals who **would** commit a new offence if prosecuted. To see this, compare the pre-reform prosecution rates for defendants who **would not** commit an offence if prosecuted (first two bounds in Panel (a)) to the pre-reform prosecution rates for those who **would** commit an offence if prosecuted (first two bounds in Panel (b)). 93%–100% of defendants of either race who would **would not** commit a new offence if prosecuted are being prosecuted, while the same is true for 75.1%–89.8% of defendants who **would** commit a new offence if prosecuted. This behavior is true in the post-reform period as well. After the reform, 74.7%–83.6% of defendants of either race who would **would not** commit a new offence if prosecuted are being prosecuted, while the same is true for 49.9%–62.8% of defendants who **would** commit a new offence if prosecuted.

This is suggestive evidence that prosecutors are less willing to prosecute cases that would result in detrimental outcomes for defendants. To investigate this, we compare prosecution rates for defendants who would have had the same treatment effect of prosecution, following Equation 8. As discussed in Section 3, this exercise 1) requires restricting the distribution of treatment effects and 2) also bounding the treatment effect on the treated. To address 1), we assume that prosecution can either induce more crime (i.e., be criminogenic) or have no impact.

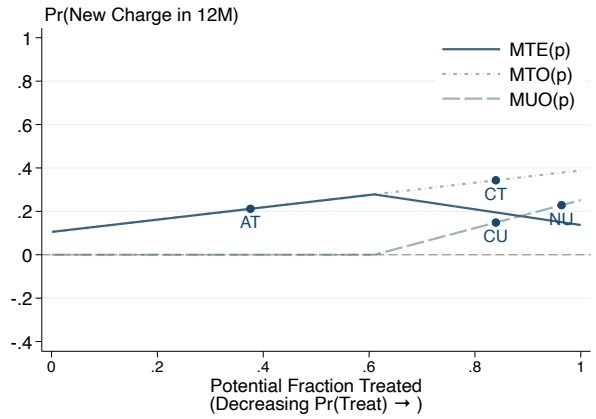
This rules out that prosecution can reduce future criminal activity, which is almost surely not satisfied at the individual-level in the population. However, violations of this assumption are likely small in this context. The literature on the deterrence effects of punishment typically finds small effects that mostly operate through incapacitation (e.g., jail), which is uncommon in our context of misdemeanor cases in Washington (Chalfin and McCrary, 2017). To examine the distribution of treatment effects in our context, we identify the marginal treatment effect (MTE) function under an auxiliary assumption of linearity in the relationship between likelihood of prosecution and potential outcomes.[61] Figure B29 shows that the MTE function in our context is always positive (i.e., increasing future criminal activity), which suggests that prosecution does not reduce future criminal activity for many people in this setting. Finally, since the treatment effect of prosecution is a function of the re-offence outcomes if dismissed, we need to make the same assumption (**A3**) of parallel trends in re-offence outcomes if dismissed as in the previous subsection to account for time trends in re-offence outcomes if dismissed.

Under these assumptions, Figure B30 shows bounds of prosecution rates separately for defendants where prosecution has no impact (Panel (a)) and where prosecution would be detrimental by inducing future criminal activity (Panel (b)). The bounds are much wider here, given the additional objects that we have to bound here. Before the reform, 90.3% – 98.1% of cases where prosecution would have no impact on future criminal activity were being prosecuted, while the same is true for 59.2%– 96.4% of cases where prosecution would be harmful. After the reform, 69.9% – 81.9% of cases where prosecution would have no impact on future criminal activity were being prosecuted,

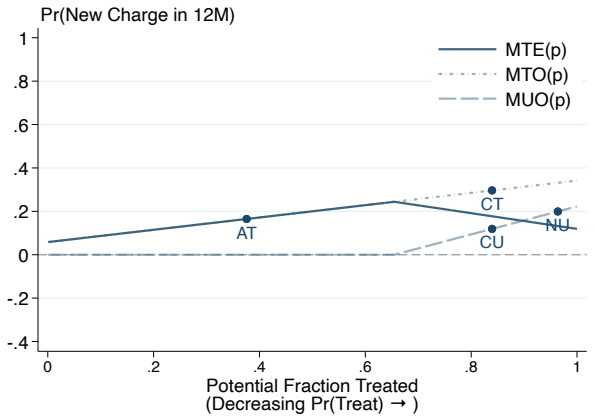
[61] As described in the main text, this assumption results in values for the marginal treated and untreated outcomes that lie outside the support of the potential outcomes (i.e., between 0 and 1). For this exercise, we constrain values of the marginal treated and untreated outcomes that are negative to be zero.

Figure B29: Marginal treatment response functions for prosecution

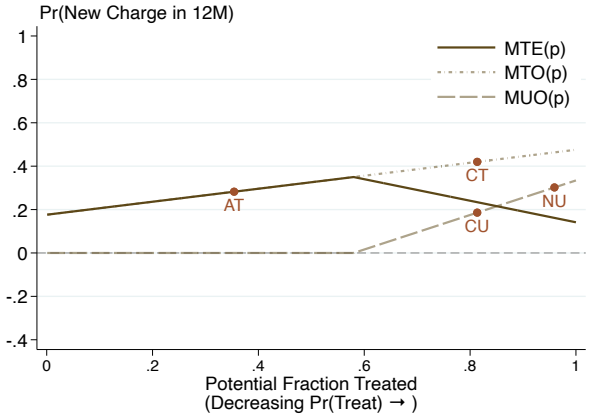
(a) White defendants, pre-reform ($Z = 0$)



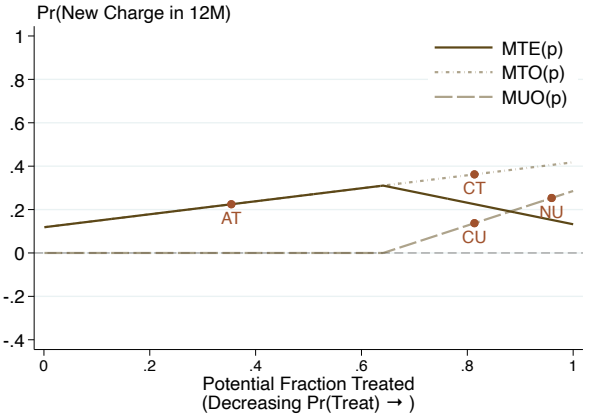
(b) White defendants, post-reform ($Z = 1$)



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



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



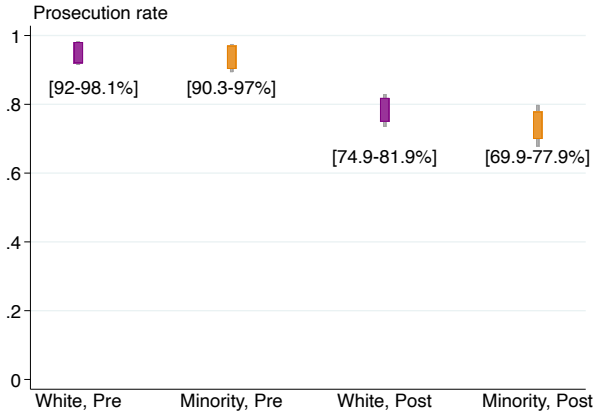
Note: This figure displays the marginal treatment response functions (Mogstad, Santos, and Torgovitsky, 2018) for prosecution. The treatment is prosecution and the outcome is whether an individual is charged with a new offence within one year after disposition. $MTO(p)$, $MUO(p)$ and $MTE(p)$ represent the marginal treated outcome, marginal untreated outcome, and marginal treatment effect functions respectively. These are identified by assuming a linear relationship between potential outcomes of always takers ('AT'), compliers ('CT'), and never takers ('NT') and their treatment propensities. Since $MTO(p)$ and $MUO(p)$ should lie within the support of the outcome (i.e., between 0 and 1), we constrain values of $MTO(p)$ and $MUO(p)$ that are negative to be zero. Lower values of the x-axis denote individuals who are more likely to be prosecuted.

while the same is true for 27.5%– 76.6% of cases where prosecution would be harmful.

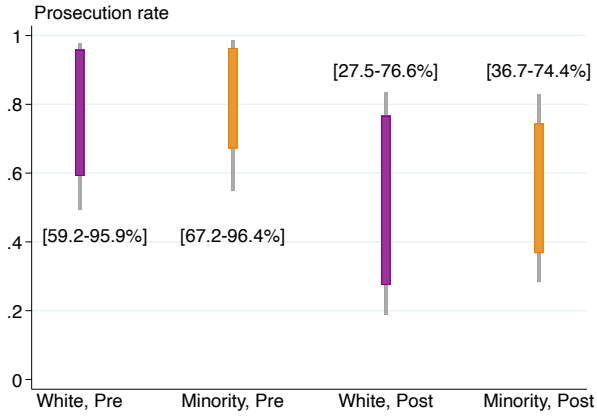
Since these bounds overlap, we cannot rule out that the prosecution rates are the same between defendants who would and would not be harmed by prosecution. However, given that the region of overlap is small, we interpret this as suggestive evidence that prosecutors may have been less likely to prosecute cases where it is likely to be criminogenic. This type of targeting behavior is consistent with recent work showing that prosecutors have some information about potential outcomes of cases and may choose not to prosecute defendants who may have higher risk of future contact with the criminal legal system (Agan, Doleac, and Harvey, 2023; Harrington, Murdock III, and Shaffer, 2023).

Figure B30: Prosecution rates conditional on treatment effect of prosecution (π_{zry})

(a) No impact, $Y_{it}(1) - Y_{it}(0) = 0$



(b) Criminogenic impact, $Y_{it}(1) - Y_{it}(0) = 1$



Note: This figure presents bounds on the average prosecution rates for each race group and time period, conditional on the treatment effect of prosecution using the approach described in Section 3. The outcome is whether an individual is charged with a new offence within one year after disposition. Confidence intervals are for the true parameter and are bootstrapped using 1,000 replications and a Bayesian bootstrap.

Appendix C Methodological Details

C.1 Comparing change in $\triangle_{\mathbf{z}}$ to observed change in discrimination

Denote the change in discrimination due to a reform by $\triangle_{\text{change}} = \triangle_1 - \triangle_0$, following the definition of $\triangle_z$ in Definition 2. For simplicity, consider that the potential outcomes are binary. The observational analogue to quantifying how an intervention affected group gaps in treatment might be to estimate the group gap in treatment responses to the intervention, defined in Definition 3: e.g., by estimating the following regressions:

$$D_i = \alpha_w + \beta_w Z_i + \varepsilon_{iw}, \text{ if } R_i = w$$
$$D_i = \alpha_m + \beta_m Z_i + \varepsilon_{ib}, \text{ if } R_i = m$$

Definition 3. Observed change in group treatment gaps due to policy reform ($\triangle_{\text{obs}}$)

$$\triangle_{\text{obs}} \equiv \beta_w - \beta_m, \text{ where } \beta_r = E[D_i|Z=1, R_i=r] - E[D_i|Z=0, R_i=r] \tag{11}$$

We are interested in whether the observed change in treatment responses coincides with the change in discrimination among individuals with identical treated potential outcomes. First, note that each expectation in Definition 3 is a weighted average of treatment rates for individuals with each potential outcome, where the weights are the prevalence of each binary value of the treated outcome. Equation 12 shows this, where $\pi_{zry} \equiv E[D_i|Z=z, R_i=r, Y_i(1)=y]$. The third and fourth lines follow from the fact that $Y_i(1) \perp Z$, since Z is random.

$$\begin{aligned} E[D_i|Z=z, R_i=r] &= E[Y_i(1)=1|Z=z, R_i=r]E[D_i|Z=z, R_i=r, Y_i(1)=1] \\ &\quad + E[Y_i(1)=0|Z=z, R_i=r]E[D_i|Z=z, R_i=r, Y_i(1)=0] \\ &= E[Y_i(1)=1|R_i=r]E[D_i|Z=z, R_i=r, Y_i(1)=1] \\ &\quad + E[Y_i(1)=0|R_i=r]E[D_i|Z=z, R_i=r, Y_i(1)=0] \\ \implies E[D_i|Z=z, R_i=r] &= Pr(Y_i(1)=1|R_i=r)\pi_{zr1} + (1 - Pr(Y_i(1)=1|R_i=r))\pi_{zr0} \end{aligned} \tag{12}$$

Plugging this into Definition 3, we can rewrite the observed change ($\triangle_{\text{obs}}$) in terms of these group- and outcome-specific treatment averages:

$$\begin{aligned} \triangle_{\text{obs}} = &[(\mu_w\pi_{1w1} + (1-\mu_w)\pi_{1w0}) - (\mu_w\pi_{0w1} + (1-\mu_w)\pi_{0w0})] \\ &- [(\mu_m\pi_{1m1} + (1-\mu_m)\pi_{1m0}) - (\mu_m\pi_{0m1} + (1-\mu_m)\pi_{0m0})] \end{aligned} \tag{13}$$

Similarly, Equation 14 rewrites the change in discrimination as a function of group- and outcome-specific treatment averages:

$$
\begin{aligned}
\triangle_{\text{change}} =& Pr(Y_i(1)=1)[\underbrace{\triangle_{11}-\triangle_{01}}_{\text{Change, } Y_i(1)=1}] + (1-Pr(Y_i(1)=1))[\underbrace{\triangle_{10}-\triangle_{00}}_{\text{Change, } Y_i(1)=0}] \\
=& \underbrace{[Pr(Y_i(1)=1)(\pi_{1w1}-\pi_{1m1}) + (1-Pr(Y_i(1)=1))(\pi_{1w0}-\pi_{1m0})]}_{\text{Gap when } Z=1} \\
&- \underbrace{[Pr(Y_i(1)=1)(\pi_{0w1}-\pi_{0m1}) + (1-Pr(Y_i(1)=1))(\pi_{0w0}-\pi_{0m0})]}_{\text{Gap when } Z=0}
\end{aligned} \tag{14}
$$

Comparing Equation 13 and Equation 14, there are only 2 cases in which $\triangle_{\text{change}} = \triangle_{\text{obs}}$:

Case 1. $Y_i(1)$ is similar across groups: $Cov(Y_i(1), R_i) = 0 \implies Pr(Y_i(1) = 1|R_i = r) = Pr(Y_i(1) = 1)\ \forall r \in R_i$

Case 2. Group-specific policy-induced responses are constant across $Y_i(1)$: $\pi_{1r1} - \pi_{0r1} = \pi_{1r0} - \pi_{0r0} = \theta_r$

The following shows how $\triangle_{\text{change}} = \triangle_{\text{obs}}$ under the described conditions. We will consider each case in turn.

Case 1

Consider that groups are similar on unobservables. Then, substituting $Pr(Y_i(1) = 1)$ for each $Pr(Y_i(1) = 1|R_i = r)$ in Equation 13, we have:

$$
\begin{aligned}
\triangle_{\text{obs}} =& [(Pr(Y_i(1)=1)\pi_{1w1} + (1-Pr(Y_i(1)=1))\pi_{1w0}) - (Pr(Y_i(1)=1)\pi_{0w1} + (1-Pr(Y_i(1)=1))\pi_{0w0})] \\
&- [(Pr(Y_i(1)=1)\pi_{1m1} + (1-Pr(Y_i(1)=1))\pi_{1m0}) - (Pr(Y_i(1)=1)\pi_{0m1} + (1-Pr(Y_i(1)=1))\pi_{0m0})] \\
=& Pr(Y_i(1)=1)[(\pi_{1w1}-\pi_{0w1}) - (\pi_{1m1}-\pi_{0m1})] + (1-Pr(Y_i(1)=1))[(\pi_{1w0}-\pi_{0w0}) - (\pi_{1m0}-\pi_{0m0})] \\
=& \triangle_{\text{change}}
\end{aligned}
$$

Case 2

Allow groups to differ on unobservables. However, assume that the policy-induced treatment response for each group is constant across constant across $Y_i(1)$: $\pi_{1r1} - \pi_{0r1} = \pi_{1r0} - \pi_{0r0} = \theta_r$. Reorganizing Equation 13 to group the policy-induced treatment response terms by race and potential outcome level:

$$
\begin{aligned}
\triangle_{\text{obs}} =& [Pr(Y_i(1)=1|R_i=w)(\pi_{1w1}-\pi_{0w1}) + (1-Pr(Y_i(1)=1|R_i=w))(\pi_{1w0}-\pi_{0w0})] \\
&- [Pr(Y_i(1)=1|R_i=m)(\pi_{1m1}-\pi_{0m1}) + (1-Pr(Y_i(1)=1|R_i=m))(\pi_{1m0}-\pi_{0m0})] \\
=& \underbrace{\theta_w - \theta_m}_{\text{Gap in potential outcome-invariant responses}}
\end{aligned}
$$

Substituting θ_r for $\pi_{1ry} - \pi_{0ry}$ in Equation 14, we see that $\triangle_{\text{change}} = \triangle_{\text{obs}}$.

$$\begin{aligned}\triangle &= Pr(Y_i(1) = 1)[\theta_w - \theta_m] + (1 - Pr(Y_i(1) = 1))[\theta_w - \theta_m] \\ &= \theta_w - \theta_m = \triangle_{\text{obs}}\end{aligned}$$

C.2 Point identifying discrimination

We first sketch a simple model of selection. Individuals are treated if the benefit of treatment, $p_i(Y_i(1), Y_i(0), Z)$, outweighs the cost. Let Z be 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 $p_i(Y_i(1), Y_i(0), Z) \in [0, 1]$. $u_i \in U[0, 1]$ is a unidimensional measure summarizing possibly multiple factors that determine an individual's cost of treatment. Given the IV assumptions listed in Section 2, $Y_i(D_i)$ 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 3, assume that 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).[62] 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 follows Section 3, except that we estimate average treated or untreated potential outcomes by integrating $MTO(p)$ or $MUO(p)$ over the full support of the treatment propensity. As a result, we obtain point estimates for the average treated/untreated outcomes and discrimination estimands.

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

We briefly demonstrate point identifying discrimination using the context of racial discrimination in incarceration decisions with 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

[62] Vytlacil (2002) demonstrates how latent index selection models coincide with the potential outcomes framework of Imbens and Angrist (1994).

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.[63]

Figure C1 validates this natural experiment separately for white and minority defendants. Panels a) and b) show that the reform resulted in a 6pp (7%) increase in non-incarceration among white defendants and a 14pp (20%) increase among minority defendants. Panels c) and d) demonstrate that future criminal activity falls by 3.1pp for white defendants (–32.6%) and increases by 3.8pp 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 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 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.[64]

Figure C2 plots the estimated race-specific marginal treated outcome functions, $MTO_r(p)$ and Table C1 integrates these functions to estimate the average outcome that would be realized if everyone was released. The point estimates suggest meaningful differences in underlying potential outcomes. 9.7% of white defendants would re-offend if released, while 17.7% of minority defendants would, although these difference are not significant.

Table C1: Average re-offence estimates (p.p.)

	Average (1)	White (2)	Minority (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 3, separately by race. The treatment is non-incarceration and the treated outcome is whether an individual is charged with a new offence 12M after disposition. Confidence intervals are bootstrapped using 1,000 replications and a Bayesian bootstrap.

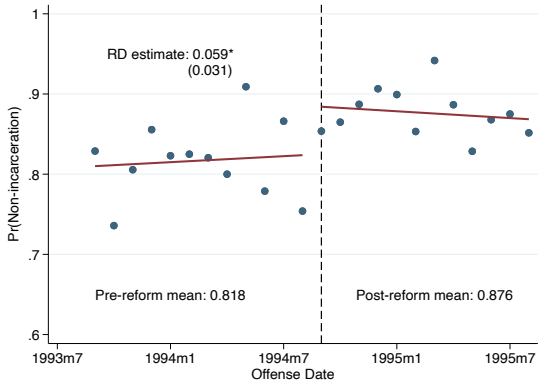
Following, Equation 3, Table C2 displays point estimates of racial discrimination in non-incarceration decisions that condition on re-offence outcomes if not incarcerated. Prior to the reform, white individuals were 12.3pp more likely to be released than minority individuals and the reform significantly narrowed this disparity. After the reform, release rates were 3.4pp higher for White defendants, a 8.9pp reduction.

[63] 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.

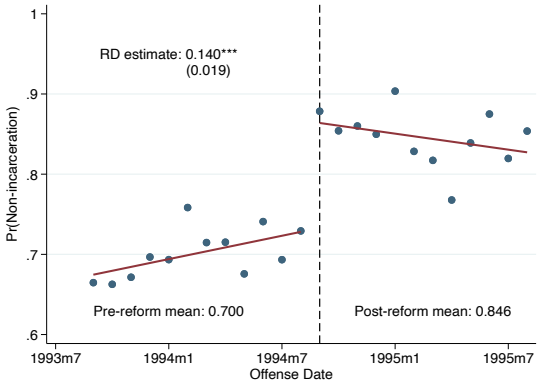
[64] We alternatively could have followed the implementation in Section 4. However, given the general lack of trends in treatment and in re-offence outcomes here, the implementation are unlikely to meaningfully differ.

Figure C1: Validating felony reform

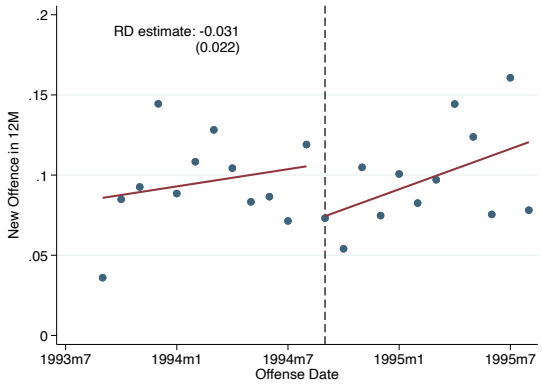
(a) Pr(Non-incarceration), white



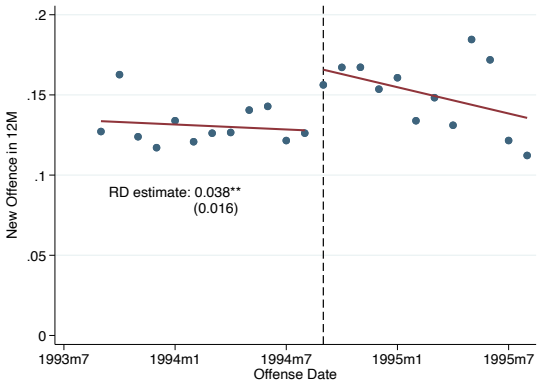
(b) Pr(Non-incarceration), minority



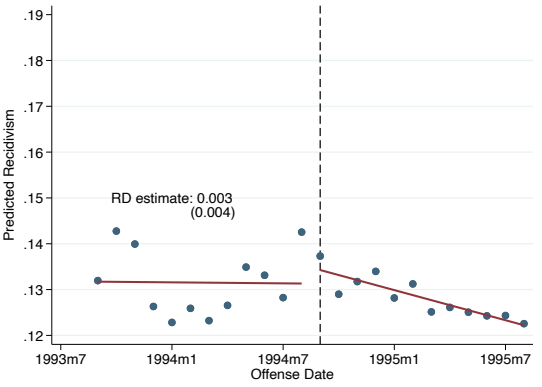
(c) Pr(New offence in 12M), white



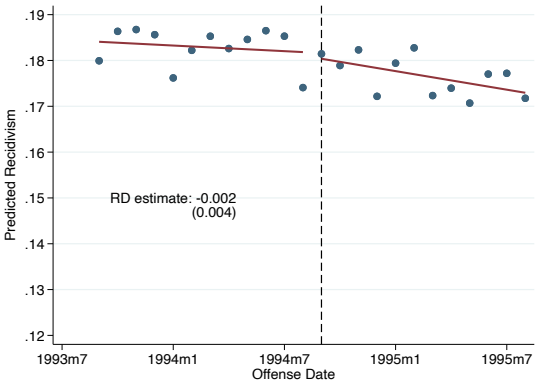
(d) Pr(New offence in 12M), minority



(e) Predicted 12M Recidivism, white

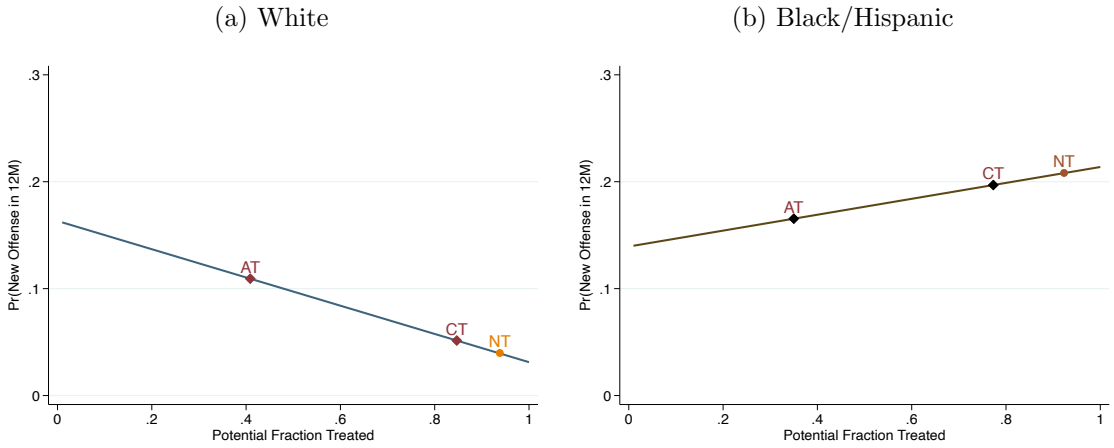


(f) Predicted 12M Recidivism, minority



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.

Figure C2: Average re-offence extrapolation by race



Note: Marginal treated outcome curves identified using $Z = t \geq \mathbf{1}$(September 1, 1994) as a binary instrument. The outcome used is whether an individual commits a new offence within 1 year, if not incarcerated. "Potential fraction treated" refers to the fraction of the population that is treated. Lower values denote individuals who are more likely to be released. As we move from 0 to 1, the fraction increases as individuals relatively less likely to be released are released. 'AT', 'CT', and 'NT' denote the estimated treated outcomes for always takers compliers and never takers respectively. The diamonds reflect outcomes of the median individual in that group.

Table C2: 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 3. The treatment is non-incarceration and the treated outcome, denoted by $Y_i(1)$, is whether an individual is charged with a new offence 12M after disposition. Confidence intervals are bootstrapped using 1,000 replications and a Bayesian bootstrap.

C.3 Identifying average potential outcomes with difference-in-difference designs

This section describes conditions under which we can non-parametrically bound average potential outcomes and average treatment effects in DiD settings with individual-level treatment non-compliance and heterogeneity in potential outcomes. The approach involves viewing the policy adoption as randomly assigned within the affected county. However, 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 LATE from recent work (De Chaisemartin and D'Haultfœuille, 2018). However, we require stronger assumptions to identify average potential outcomes and treatment effects 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.[65]
- $Y_{it}(d,g)$: Potential outcomes, given the time period, their treatment state 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 typically assumed 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). However, 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 to 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:[66]

[65] Note that this allows individuals in either county to be treated both before and after the policy, unlike county-level treatment in typical DiD.

[66] 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. This is enough to identify the LATE, but does not allow us to identify the average potential outcomes of each compliance group separately. This is because 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.

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.

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

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 this setup, 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 $G = 1$.

$$\begin{aligned} p_A &= E[D_i|G = 1, T = 0] \\ p_N &= 1 - (E[D_i|G = 1, T = 1]) \\ p_C &= 1 - (p_A + p_N) \end{aligned} \tag{16}$$

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 17 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 17).

$$\begin{aligned} E[Y_i|D_i = 1, G = 1, Z = 0] &= E[Y_i|D_i = 1, G = 1, T = 0] = E[Y_{i0}(1,1)|A, G = 1] \\ \underbrace{E[Y_{i1}(1,1)|A, G = 1]}_{\text{Unobserved}} &= \underbrace{E[Y_{i0}(1,1)|A, G = 1]}_{\text{Observed}} + \underbrace{\theta_1}_{\text{Unobserved}} \end{aligned} \tag{17}$$

Equation 18 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 18).

$$\begin{aligned} E[Y_i|D_i=0,G=1,Z=1] &= E[Y_i|D_i=0,G=1,T=1] = E[Y_{i1}(0,1)|N,G=1] \\ \underbrace{E[Y_{i0}(0,1)|N,G=1]}_{\text{Unobserved}} &= \underbrace{E[Y_{i1}(0,1)|N,G=1]}_{\text{Observed}} - \underbrace{\theta_0}_{\text{Unobserved}} \end{aligned} \tag{18}$$

We can use Assumption 5 to 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. Hence the average change in treated outcomes in $G=0$ identifies the time trend for always takers in $G=1$, since Assumption 5 states that time trend in potential treated outcomes for always takers are identical across counties and is equal to the time trend for compliers (see Equation 19).

$$\begin{aligned} \theta_1 = E[Y_{i1}(1,1) - Y_{i0}(1,1)|G=1,A] &= E[Y_{i1}(1,1) - Y_{i0}(1,1)|G=1,C] \\ &= E[Y_{i1}(1,0) - Y_{i0}(1,0)|G=0] \end{aligned} \tag{19}$$

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 20).

$$\begin{aligned} \theta_0 = E[Y_{i1}(0,1) - Y_{i0}(0,1)|G=1,N] &= E[Y_{i1}(0,1) - Y_{i0}(0,1)|G=1,C] \\ &= E[Y_{i1}(0,0) - Y_{i0}(0,0)|G=0] \end{aligned} \tag{20}$$

Equation 21 restates this by combining this with Equations 17 and 18, 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, $G=0$.

$$
\begin{aligned}
E[Y_{i1}(1,1)|A,G=1] &= E[Y_{i0}(1,1)|A,G=1] + \theta_1 \\
&= E[Y_{i0}(1,1)|A,G=1] + E[Y_{i1}(1,0) - Y_0(1,0)|G=0] \\
E[Y_{i0}(0,1)|N,G=1] &= E[Y_{i1}(0,1)|N,G=1] - \theta_0 \\
&= E[Y_{i1}(0,1)|N,G=1] - E[Y_{i1}(0,0) - Y_{i0}(0,0)|G=0]
\end{aligned}
\tag{21}
$$

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 22 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 21, $E[Y_i|D_i=1,G=1,T=1]$ is a sample average, and each of the proportions is obtained using Equation 16.

$$
\begin{aligned}
E[Y_i|D_i=1,G=1,T=1] &= \frac{p_A E[Y_{i1}(1,1)|A,G=1] + p_C E[Y_{i1}(1,1)|C,G=1]}{p_A + p_C} \\
E[Y_{i1}(1,1)|C,G=1] &= \frac{(p_A+p_C)E[Y_i|D_i=1,G=1,T=1] - p_A E[Y_{i1}(1,1)|A,G=1]}{p_C}
\end{aligned}
\tag{22}
$$

We can estimate untreated outcomes for compliers in a similar way. The first line of Equation 23 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 21, $E[Y|D_i=0,G=1,T=0]$ is a sample average, and each of the proportions is obtained using Equation 16.

$$
\begin{aligned}
E[Y_i|D_i=0,G=1,T=0] &= \frac{p_N E[Y_{i0}(0,1)|N,G=1] + p_C E[Y_{i0}(0,1)|C,G=1]}{p_N + p_C} \\
E[Y_{i0}(0,1)|C,G=1] &= \frac{(p_N+p_C)E[Y_i|D_i=0,G=1,T=0] - p_N E[Y_{i0}(0,1)|N,G=1]}{p_C}
\end{aligned}
\tag{23}
$$

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, $G = 0$.

$$\begin{aligned} E[Y_{i0}(1,1)|C, G = 1] &= E[Y_{i1}(1,1)|C, G = 1] - \theta_1 \\ E[Y_{i1}(0,1)|C, G = 1] &= E[Y_{i0}(0,1)|C, G = 1] + \theta_0 \end{aligned} \tag{24}$$

We now have average treated and untreated outcomes for each compliance group and in each period.

C.4 Bounds when conditioning on $\mathbf{Y_i(0)}$ or $\mathbf{Y_i(1) - Y_i(0)}$

Conditioning on $Y_i(0)$

Equation 25 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 2. The third line uses the fact that the average untreated outcome observed if everyone was 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 7.

$$\begin{aligned} 1 - E[D_i = 0|R_i = r, Y_i(0) = y] &= 1 - \frac{E[Y_i(0) = y|R_i = r, D_i = 0] \times (1 - \bar{D})}{E[Y_i(0) = y|R_i = r]} \\ &= \frac{E[Y_i(0) = y|R_i = r] - E[Y_i(0) = y|R_i = r, D_i = 0] \times (1 - \bar{D})}{E[Y_i(0) = y|R_i = r]} \\ &= \frac{E[Y_i(0) = y|R_i = r, D_i = 1] \times \bar{D}}{E[Y_i(0) = y|R_i = r]} \\ &= E[D_i|R_i = r, Y_i(0) = y] \end{aligned} \tag{25}$$

Conditioning on $Y_i(1) - Y_i(0)$

The equation below reproduces Equation 8, where $\tau_i \equiv Y_i(1) - Y_i(0)$. The right hand side is a function of two partially identified objects (first term in numerator and denominator) and a point estimate (second term in numerator). The two partially identified objects are dependent – the denominator is a function of the first term in the numerator.

$$E[D_i|Z=z, R_i=r, \tau_i=y] = \frac{E[\tau_i=y|Z=z, R_i=r, D_i=1] \times E[D_i|Z=z, R_i=r]}{E[\tau_i=y|R_i=r]}$$

We rewrite the treatment rate conditional on the treatment effect, omitting group conditioning for brevity and considering the case where the policy reform has taken effect, i.e. $Z=1$. Here, we assume that $\tau_i \in \{0,1\}$, because this assumption is required to identify this type of treatment rate, as described in Section 3.

$$\begin{aligned} E[D_i|Z=1, \tau_i=1] &= \frac{E[\tau_i|Z=z, D_i=1] \times E[D_i|Z=z]}{E[\tau_i]} \\ &= \frac{\left(p_A E[\tau_i|\text{A}] + p_C E[\tau_i|\text{C}]\right)/(p_A+p_C)}{p_A E[\tau_i|\text{A}] + p_C E[\tau_i|\text{C}] + p_N E[\tau_i|\text{N}]} \times E[D_i|Z=1] \end{aligned} \tag{26}$$

The average treatment effect on the treated is a function of the bounds on the treatment effect for always takers and the point estimated treatment effect for compliers. The denominator ($E[\tau_i]$) is a function of two partially identified objects, $E[\tau_i|\text{A}]$ and $E[\tau_i|\text{N}]$, since we never observe untreated outcomes for always takers or treated outcomes for never takers.

This implies that the treatment rate conditional on the treatment effect is of the form $y = f(p,q) = \frac{p \times r}{p+q}$, where r is a point estimate and the other quantities are bounds. We compute bounds for y by holding one partially identified object fixed at a time, evaluating the function at the extremes of the other partially identified object. That is, we take the minimum and maximum values of y over the following cases: $f(\underline{p}, \underline{q})$, $f(\underline{p}, \overline{q})$, $f(\overline{p}, \underline{q})$, $f(\overline{p}, \overline{q})$.

C.5 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 race are overlapping. To see this, consider the following three cases:[67]

[67] 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.

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), enforcing the weak monotonicity restriction implied by each bootstrap replication. For each 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.

C.6 Extrapolation discrimination using averages identified at RD cut-off

We can apply the average treated outcome estimates derived from information at the cut-off to adjust treatment rates in portions of the analysis sample that are away from the cut-off under certain assumptions. 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. Finally, any assumptions discussed below will need to hold by subgroup.

Assume that the following hold (omitting subgroup notation here):

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

Let us consider these assumption in the context of the application in Section 4. RD1 states that within the analysis window, the test score does not influence promotion decisions by itself – only the cut-off does. RD1 also assumes that the size of the first-stage would be the same in counterfactuals where the test score cut-off was placed 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. These assumptions are stronger than those in typical approaches to extrapolate away from RD cut-offs. This is because the focus here is on extrapolating average potential outcomes rather than treatment effects (Cattaneo et al., 2021; Ricks, 2022).

In our setting, we fail to reject that running variable alone influences treatment rates, which is suggestive evidence that RD1 is not violated here. To assess whether RD2 is reasonable, note that we can split $E[Y_i(1)|s = s^*]$ into average treated outcomes for always takers, compliers and

never takers. 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 27 must hold.

$$
\begin{aligned}
&E[Y_i(1)|s = s^*] = E[Y_i(1)|s<s^*] \\
&\implies p_A E[Y_i(1)|A, s = s^*] + p_C E[Y_i(1)|C, s = s^*] + p_N E[Y_i(1)|N, s = s^*] = \\
&\qquad p_A E[Y_i(1)|A, s<s^*] + p_C E[Y_i(1)|C, s<s^*] + p_N E[Y_i(1)|N, s<s^*] \\
&\implies p_C \left(E[Y_i(1)|C, s = s^*] - E[Y_i(1)|C, s<s^*]\right) + p_N \left(E[Y_i(1)|N, s = s^*] - E[Y_i(1)|N, s<s^*]\right) = \\
&\qquad = p_A \left(E[Y_i(1)|A, s<s^*] - E[Y_i(1)|A, s = s^*]\right)
\end{aligned}
\tag{27}
$$

Whether this is a reasonable assumption is an empirical question. In Figure 3, 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 final expression of Equation 27 is negative. The only way for this assumption to hold with equality is for the left hand side 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.

If we are in a situation where these assumptions might be violated, it is useful to consider the direction of bias that arises from applying average potential outcomes identified at an RD cut-off to the entire sample. We introduce additional notation, assuming the potential outcomes are binary:

- True & estimated average treated outcome by group: $E[Y_i(1)|R_i = r] = \mu_r$ and $\widehat{\mu}_r$
- True & estimated average treated outcome in population: $E[Y_i(1)] = \mu_r$ and $\widehat{\widehat{\mu}}$
- $R_i \in \{h, l\}$

Following Equation 3, we derive the bias in the period-, group- and outcome-specific treatment rates:

$$
\begin{aligned}
\widehat{\pi}_{zr1} &= \pi_{zr1}\left(\frac{\mu_r}{\widehat{\mu}_r}\right) \\
\widehat{\pi}_{zr0} &= \pi_{zr0}\left(\frac{1-\mu_r}{1-\widehat{\mu}_r}\right) \\
\widehat{\bar{\mu}} - \bar{\mu} &= p_h(\widehat{\mu}_h - \mu_h) + p_l(\widehat{\mu}_l - \mu_l)
\end{aligned} \tag{28}
$$

Since μ_r and $1-\mu_r$ are always positive (they represent shares of individuals with a given value of $Y_i(D_i)$), we sign the bias that results from biased estimates of μ_r. Consider that $\widehat{\mu}_r > \mu_r$, which would be the case in the example above that extrapolates μ_r from the cut-off to below the cut-off. Equation 28 implies that i) $\widehat{\pi}_{zr1} < \pi_{zr1}$, ii) $\widehat{\pi}_{zr0} > \pi_{zr0}$. That is, applying information from the cut-off to below the cut-off in such a situation would lead one to i) underestimate the promotion rate of students who are ready for 4th grade and ii) overestimate the promotion rate of students who are not ready. If $\widehat{\mu}_r > \mu_r$ for both groups, then iii) $\widehat{\bar{\mu}} > \bar{\mu}$, i.e., one would overestimate the overall readiness for 4th grade in the population.

However, the implications for potential outcome-specific gaps below/above the cut-off ($\triangle_{zy}$) and average gaps below/above the cut-off ($\triangle_z$) are ambiguous. Recall that

$$
\begin{aligned}
\triangle_{zy} &= \pi_{zhy} - \pi_{zly} \\
\triangle_z &= \bar{\mu}\,\triangle_{z1} + (1-\bar{\mu})\triangle_{z0}
\end{aligned}
$$

Even if the estimates for π_{zhy} and π_{zly} are both biased in the same direction, the relative magnitudes of the bias in π_{zry} for each group will determine the overall bias in $\triangle_{zy}$. Similarly, the impact on $\triangle_z$ is ambiguous since the estimates of $\triangle_{z1}$ and $\triangle_{z0}$ will be biased in different directions.