NBER WORKING PAPER SERIES

MEASURING RACIAL DISCRIMINATION IN BAIL DECISIONS

David Arnold
Will S. Dobbie
Peter Hull

Working Paper 26999
http://www.nber.org/papers/w26999

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
April 2020

We thank Josh Angrist, Tim Armstrong, Leah Boustan, Sydnee Caldwell, Raj Chetty, John Donohue, Joseph Doyle, Matt Gentzkow, Ed Glaeser, Paul Goldsmith-Pinkham, Felipe Goncalves, Damon Jones, Conrad Miller, Derek Neal, Scott Nelson, Sam Norris, Jesse Shapiro, Megan Stevenson, Crystal Yang, and numerous seminar participants for helpful comments and suggestions. Emily Battaglia, Nicole Gandre, Jared Grogan, Ashley Litwin, Alexia Olaizola, Bailey Palmer, Elise Parrish, Emma Rackstraw, and James Reeves provided excellent research assistance. The data we analyze are provided by the New York State Division of Criminal Justice Services (DCJS), and the Oÿce of Court Administration (OCA). The opinions, findings, and conclusions expressed in this publication are those of the authors and not those of DCJS, OCA, or the National Bureau of Economic Research. Neither New York State, DCJS or OCA assumes liability for its contents or use thereof.

Measuring Racial Discrimination in Bail Decisions
David Arnold, Will S. Dobbie, and Peter Hull
NBER Working Paper No. 26999
April 2020, Revised in October 2020
JEL No. C26,J15,K42

ABSTRACT

We develop new quasi-experimental tools to measure racial discrimination, due to either racial bias or statistical discrimination, in the context of bail decisions. We show that the omitted variables bias in observational release rate comparisons can be purged by using the quasi-random assignment of judges to estimate average race-specific misconduct risk. We find that approximately two-thirds of the average release rate disparity between white and Black defendants in New York City is due to racial discrimination. We then develop a hierarchical marginal treatment effects model to study the drivers of discrimination, finding evidence of both racial bias and statistical discrimination. Outcome-based tests of racial bias therefore omit an important source of racial discrimination in bail decisions, and cannot be used to rule out all possible violations of U.S. anti-discrimination law.

David Arnold
9500 Gilman Drive
ECON 229
UCSD
United States
San Diego, CA 92093
daarnold@ucsd.edu

Peter Hull
University of Chicago
5757 South University Avenue
Chicago, IL 60637
and NBER
hull@uchicago.edu

Will S. Dobbie
Harvard Kennedy School
79 John F. Kennedy St.
Cambridge, MA 02138
and NBER
will_dobbie@hks.harvard.edu

1 Introduction

Racial disparities are pervasive throughout much of the U.S. criminal justice system. Compared to observably similar white individuals, Black individuals are more likely to be searched by the police, charged with a serious crime, detained before trial, convicted of an offense, and incarcerated.[1] Such racial disparities are often taken as evidence of discrimination driven by racially biased preferences or stereotypes. But this interpretation overlooks two alternative explanations. First, the observed disparities may reflect legally relevant differences in criminal behavior that are partially observed by police officers, prosecutors, and judges but not by the econometrician. Second, the observed disparities may also reflect discrimination driven by statistical discrimination, not just racially biased preferences and stereotypes. Distinguishing between these different explanations and correctly measuring racial discrimination remains difficult, hampering efforts to formulate appropriate policy responses.

This paper develops new quasi-experimental tools to measure racial discrimination, regardless of its source. We study bail decisions, where the sole legal objective of judges is to allow most defendants to be released before trial while minimizing the risk of pretrial misconduct (such as failing to appear in court or being arrested for a new crime). Bail judges thus risk violating U.S. anti-discrimination law if they release white and Black defendants with the same objective misconduct potential at different rates. Correspondingly, we measure discrimination as the difference in a judge's release rates between white and Black individuals with identical misconduct potential. This measure is consistent with mainstream legal views on what constitutes discrimination in the criminal justice system (Yang and Dobbie, 2019), as well as economic notions of discrimination that compare the treatment of white and Black individuals with the same productivity (Aigner and Cain, 1977) and notions of algorithmic discrimination in the computer science literature (Berk et al., 2018). Importantly, our measure captures discrimination arising from either racial bias or statistical discrimination. Since this measure can be understood as isolating each judge's legally unwarranted release rate disparity, we use the terms racial discrimination and unwarranted disparity interchangeably.

Estimating legally unwarranted release rate disparities among white and Black defendants is fundamentally challenging. Observational comparisons cannot control for unobserved misconduct potential and can therefore suffer from omitted variables bias (OVB) when there are unobserved racial differences in misconduct risk. Randomized audit studies (e.g., Bertrand and Mullainathan, 2004; Ewens, Tomlin and Choon Wang, 2014) can test whether decision-makers treat fictitious white and Black individuals with the same observable characteristics in the same way, but do not capture discrimination on seemingly race-neutral characteristics and are infeasible in high-stakes and face-to-face settings such as bail decisions. Outcome-based tests leveraging standard instrumental variables (IV) methods can detect racial bias at the margin of release decisions (e.g., Arnold, Dobbie and Yang, 2018; Marx, 2018), but do not speak to the presence of statistical discrimination or measure the magnitude of unwarranted disparities. Standard IV methods also require an assumption of first-stage monotonicity (Imbens and Angrist, 1994; Heckman and Vytlacil, 2005), which here imposes a strong restriction on how judges choose which defendants to release before trial.

Our primary methodological contribution is to show that racial discrimination in bail decisions, due to either racial bias or statistical discrimination, can be measured with observational release

[1] There is a large literature documenting racial disparities in the criminal justice system. See, for example, work by Gelman, Fagan and Kiss (2007), Antonovics and Knight (2009), Anwar, Bayer and Hjalmarsson (2012), Abrams, Bertrand and Mullainathan (2012), McIntyre and Baradaran (2013), and Rehavi and Starr (2014), among many others.

rate comparisons that are rescaled using quasi-experimental estimates of average white and Black misconduct risk. The OVB in observational release rate comparisons comes from the correlation between defendant race and unobserved misconduct potential in each judge's defendant pool. When judges are as-good-as-randomly assigned, this correlation is common to all judges and is a simple function of misconduct risk (i.e., average misconduct potential) by race. We can thus use estimates of race-specific misconduct risk to rescale observational release rate comparisons in order to make released white and Black defendants comparable in terms of misconduct potential within each as-good-as-randomly assigned judge's defendant pool. The rescaled comparisons reveal the extent to which each judge releases white and Black defendants with the same objective misconduct potential at different rates, even though this potential is unobserved and cannot be directly conditioned on. The key econometric challenge is then to estimate the average misconduct risk parameters, which is difficult even when judges are as-good-as-randomly assigned since misconduct outcomes are only observed for the subset of defendants who are endogenously released before trial.

We estimate the average misconduct risk parameters needed for our discrimination measure from quasi-experimental variation in pretrial release and misconduct rates, without imposing a model of judge behavior or a first-stage monotonicity assumption. To build intuition for our approach, consider a setting with a supremely lenient and as-good-as-randomly assigned bail judge who releases nearly all defendants assigned to her. The supremely lenient judge's release rates among white and Black defendants are close to one, meaning (by as-good-as-random assignment) that the misconduct rates among her released white and Black defendants are close to the average misconduct risk parameters needed for our discrimination measure. Without such a supremely lenient judge, the required average misconduct risk inputs can be estimated using model-based or statistical extrapolations of pretrial release and misconduct rates across quasi-randomly assigned judges. This extrapolation of quasi-experimental moments is analogous to a standard regression discontinuity approach of extrapolating average potential outcomes to a treatment cutoff from nearby observations. Importantly, neither our approach to extrapolating average potential outcomes nor to estimating discrimination from these extrapolations require a model of judge decision-making, only that the extrapolations of pretrial release and misconduct rates and the judges' legal objective are well-specified by the econometrician.

We use our quasi-experimental approach to measure racial discrimination in bail decisions in New York City (NYC), home to one of the largest pretrial systems in the country. Our most conservative estimates show that approximately two-thirds of the average release rate disparity between white and Black defendants is explained by racial discrimination (62 percent, or 4.2 percentage points out of 6.8 percentage points), with the remaining one-third explained by unobserved racial differences in misconduct risk. This finding applies to most defendant subgroups and is robust to different extrapolations of average misconduct risk, specifications of pretrial misconduct, classifications of pretrial release, and definitions of defendant race. Judge-specific estimates show that the vast majority of bail judges discriminate against Black defendants (87 percent, in our most conservative estimates), with higher levels of discrimination among more stringent judges, judges assigned a lower share of cases with Black defendants, and judges who are not newly appointed in our sample period.[2] Judge-specific estimates are also highly correlated over time, raising the possibility that discrimination across individual judges can be reliably targeted by a policymaker.

[2]We define a judge as newly appointed if he or she enters the data after our sample period begins and works three consecutive months of regular caseloads.

Our second methodological contribution is to develop a hierarchical marginal treatment effects (MTE) model that imposes additional structure on the quasi-experimental variation to investigate whether these unwarranted disparities in bail decisions are driven by racial bias or statistical discrimination. The model allows for judge- and race-specific risk preferences and signal quality. The latter allows for heterogenous race-specific predictive skill across judges, in violation of the conventional first-stage monotonicity assumption. The model implies a distribution of judge- and race-specific MTE curves that can be used to test for racial bias at the margin of release and measure racial differences in average risk or signal quality that otherwise generate statistical discrimination. We estimate the distribution of MTE curves using a tractable simulated minimum distance (SMD) procedure that matches moments of the quasi-experimental variation in pretrial release and misconduct rates across judges. We find evidence of both racial bias and statistical discrimination in NYC, with the latter coming from a higher level of average risk (that exacerbates discrimination) and less precise risk signals (that alleviates discrimination) for Black defendants. Outcome-based tests of racial bias (e.g., Arnold, Dobbie and Yang, 2018) therefore omit an important source of discrimination in NYC bail decisions, and cannot be used to rule out all possible violations of U.S. anti-discrimination law.

We conclude by using our model to investigate whether discrimination can be reliably targeted, and potentially reduced, with existing data. We suppose that judges can be subjected to race-specific release rate quotas that eliminate unwarranted racial disparities, as estimated by a policymaker. We find that targeting the most discriminatory NYC judges with a quota based on our quasi-experimental estimates can reduce the average level of discrimination by 36 percent, and that targeting all judges with such a quota can essentially eliminate discrimination, despite the noise in our estimation procedure. By comparison, targeting judges with a quota based on observational release rate disparities can lead to a small but non-zero level of discrimination against white defendants, due to OVB.

This paper adds to a recent empirical literature that uses quasi-experimental variation to test for bias and discrimination in the criminal justice system. Arnold, Dobbie and Yang (2018) use the release tendencies of quasi-randomly assigned bail judges to test for racial bias under a conventional first-stage monotonicity assumption, while Marx (2018) uses a similar approach to test for racial bias at the margin of police stops. We show how quasi-experimental judge assignment can be used to measure all forms of racial discrimination, not just racial bias, without any such behavioral assumptions. We further show how the drivers of this more comprehensive measure of discrimination can be investigated by imposing alternative structure on the quasi-experimental variation.[3] This structure allows us to translate differences in marginal outcomes into differences in average release rates, to quantify the importance of racial bias (unlike typical outcome-based tests, which can only test for its existence).

Methodologically, this paper adds to a recent literature on estimating average treatment effects (ATEs) and MTEs with multiple discrete instruments (Kowalski, 2016; Brinch, Mogstad and Wiswall, 2017; Mogstad, Santos and Torgovitsky, 2018; Hull, 2020). An important feature of our approach is that we do not impose the usual first-stage monotonicity assumption, which has received recent scrutiny both in general (Mogstad, Torgovitsky and Walters, 2019) and in the specific context of

[3]Other recent related work includes Rose (2020) and Feigenberg and Miller (2020). Rose (2020) shows that a policy reform that sharply reduced prison punishments for technical probation violations nearly eliminated the racial disparity in incarceration without significantly increasing the disparity in reoffending rates, suggesting that technical probation violations may convey less precise risk signals for Black individuals on probation. Feigenberg and Miller (2020) show that Black motorists in Texas are stopped at higher rates than white motorists without any commensurate increase in contraband hit rates, suggesting that the racial disparity in search rates is inefficient.

judge IV designs (Mueller-Smith, 2015; Frandsen, Lefgren and Leslie, 2019; Norris, 2019).[4] Our extrapolation-based solution to estimating ATEs (i.e., mean misconduct risk) without imposing monotonicity is most closely related to Hull (2020), who considers non-parametric extrapolations of quasi-experimental moments in the spirit of "identification at infinity" in sample selection models (Chamberlain, 1986; Heckman, 1990; Andrews and Schafgans, 1998). Our hierarchical solution to estimating a distribution of MTE curves without monotonicity is most closely related to the contemporaneous approach of Chan, Gentzkow and Yu (2020), who estimate a structural model of doctor preferences and skill in making pneumonia diagnoses.[5]

The remainder of the paper is organized as follows. Section 2 provides an overview of the NYC pretrial system. Section 3 outlines the conceptual framework underlying our analysis. Section 4 describes our data and documents release rate differences for Black and white defendants. Section 5 develops and implements our quasi-experimental approach to measuring racial discrimination in bail decisions. Section 6 develops and estimates our hierarchical MTE model to explore the drivers of discrimination. Section 7 conducts policy counterfactuals. Section 8 concludes.

2 Setting

We study racial discrimination in the New York City pretrial system, one of the largest in the country. The U.S. pretrial system is meant to allow most criminal defendants to be released from legal custody while minimizing the risk of pretrial misconduct. Bail judges in both NYC and the country as a whole are granted considerable discretion in determining which defendants should be released before trial, but they cannot discriminate against minorities and other protected classes even when membership in a protected class contains information about the underlying risk of criminal misconduct (Yang and Dobbie, 2019). Judges are also not meant to assess guilt or punishment when determining which individuals should be released from custody, nor are they meant to consider the political consequences of their bail decisions. Bail judges therefore risk violating U.S. anti-discrimination law if they release white and Black individuals with the same objective pretrial misconduct potential at different rates.

In NYC, bail conditions are set by a judge at an arraignment hearing held shortly after an arrest. These hearings usually last a few minutes and are held through a videoconference to the detention center. The judge typically receives detailed information on the defendant's current offense and prior criminal record, as well as a release recommendation based on a six-item checklist developed by a local nonprofit (New York City Criminal Justice Agency Inc., 2016). The judge then has several options in setting bail conditions. First, she can release defendants who show minimal risk on a promise to return for all court appearances, known broadly as release on recognizance (ROR) or release without conditions. Second, she can require defendants to post some sort of bail to be released. The judge can also send higher-risk defendants to a supervised release program as an alternative to cash bail.

[4] Skepticism of conventional monotonicity in judge IV designs is as old as the assumption itself. In their initial paper on the identification of local average treatment effects, Imbens and Angrist (1994) note that in the context of administrative screening "[monotonicity] requires that if official A accepts applicants with probability $P(0)$, and official B accepts people with probability $P(1) > P(0)$, official B must accept *any* applicant who would have been accepted by official A. This is unlikely to hold if admission is based on a number of criteria" (Example 2; p. 472).

[5] Chan, Gentzkow and Yu (2020) model doctor decisions as following a hierarchical bivariate probit with variation in the latent index correlation across doctors. By comparison, we model judges as acting on posteriors from noisy risk signals with variation in signal quality across judges. We also show how this model can be used to form posterior MTE frontiers for each judge and race, and link these MTE frontiers to racial bias and statistical discrimination.

Finally, the judge can detain defendants pending trial by denying bail altogether.[6]

We exploit three features of the pretrial system in our analysis. First, the legal objective of bail judges is both narrow and measurable among the set of released defendants for whom pretrial misconduct outcomes are observed (although not among detained defendants, for whom such outcomes are unobserved). Second, bail judges can be effectively viewed as making binary "treatment" decisions, releasing low-risk defendants (generally by ROR or setting a low cash bail amount) and detaining high-risk defendants (generally by setting a high cash bail amount). We also explore different definitions of bail decisions in our analysis, such as viewing judges as deciding between release without conditions and any cash bail amount. Third, the case assignment procedures used in most jurisdictions, including NYC, generate quasi-random variation in judge assignment for defendants arrested at the same time and place. The quasi-random variation in judge assignment, in turn, generates quasi-experimental variation in the probability that a defendant is released before trial.

There are also two differences between the NYC pretrial system and other pretrial systems around the country that are potentially relevant for our analysis. First, New York state instructs judges to only consider the risk that defendants will not appear for a required court appearance when setting bail conditions (a so-called failure to appear, or FTA), not the risk of new criminal activity as in most states. We explore robustness to this New York specific definition of pretrial misconduct in our analysis. Second, many defendants in NYC will never have bail set, either because the police gave them a desk appearance ticket that does not require an arraignment hearing or because the case was dismissed or otherwise disposed at the arraignment hearing before bail was set. However, the decision of whether or not to issue a desk appearance ticket is made before the bail judge is assigned, and cases should only be dismissed or otherwise disposed at arraignment if there is a clear legal defect in the case (Leslie and Pope, 2017). We show below that there is no relationship between the assigned bail judge and the probability that a case exits our sample due to case disposal or dismissal at arraignment, and exclude these cases from our analysis.

3 Conceptual Framework

3.1 Defining Racial Discrimination

We study racial discrimination in a setting where a set of decision-makers j make binary decisions $D_{ij} \in \{0, 1\}$ for an *iid* set of individuals i. Each decision-maker's goal is to align D_{ij} with an unobserved binary state $Y_i^* \in \{0, 1\}$. In the context of bail decisions, $D_{ij} = 1$ indicates that judge j would release defendant i if assigned to her case (with $D_{ij} = 0$ otherwise) while $Y_i^* = 1$ indicates that the defendant would subsequently fail to appear in court or be rearrested for a new crime if released (with $Y_i^* = 0$ otherwise). Each judge's objective is to release individuals without misconduct potential ($Y_i^* = 0$) and detain individuals with misconduct potential ($Y_i^* = 1$), but may differ in their predictions of which individuals are in each group.[7] We note that we define D_{ij} as the potential decision of judge j for each defendant i, setting aside for now the judge assignment process which yields actual release decisions from these latent variables.

[6] Cases such as murder, kidnapping, arson, and high-level drug possession and sale almost always result in a denial of bail, though these cases make up only about 0.8 percent of our sample. By comparison, about 70 percent of defendants in NYC are released ROR each year, nearly 30 percent are assigned cash bail or one less commonly used bail options such as insurance company bail bonds, and about 1.5 percent are sent to a supervised release program.

[7] Appendix B.1 discusses how our approach can be extended to multi-valued or continuous Y_i^*.

We measure racial discrimination, both overall and for each judge, with the average release rate disparity between white and Black individuals with identical misconduct potential. Letting $R_i \in \{w, b\}$ index the race of white and Black individuals, the level of discrimination for judge j is given by:

$$\Delta_j = E[E[D_{ij} \mid Y_i^*, R_i = w] - E[D_{ij} \mid Y_i^*, R_i = b]] \tag{1}$$

The system-wide level of discrimination is given by the case-weighted average Δ_j across all judges. The inner difference of Equation (1) compares the potential release rates of white ($R_i = w$) and Black ($R_i = b$) defendants with the same objective misconduct potential Y_i^* when assigned to judge j. The outer expectation averages this conditional release rate comparison over the distribution of Y_i^*. We say that judge j discriminates against Black defendants when $\Delta_j > 0$, that she discriminates against white defendants when $\Delta_j < 0$, and that she does not discriminate against either Black or white defendants when $\Delta_j = 0$, again recognizing that the D_{ij} capture a judge's potential release decisions. By holding the potential defendant population fixed, estimates of Δ_j can be used to calculate both the average level of racial discrimination in a bail system as well as any variation in the level of discrimination across judges. As mentioned above, we interchangeably refer to Δ_j as the level of racial discrimination for judge j and the unwarranted release rate disparity for judge j.

With binary D_{ij} and Y_i^*, the Δ_j parameters can be understood as capturing racial differences in the tendency of judge j to correctly and incorrectly classify individuals by their objective misconduct potential. We let $\delta_{jr}^T = Pr(D_{ij} = 1 \mid Y_i^* = 0, R_i = r)$ denote the probability that judge j correctly releases defendants of race r without misconduct potential (her "true negative rate" for this race) and $\delta_{jr}^F = Pr(D_{ij} = 1 \mid Y_i^* = 1, R_i = r)$ denote the probability that judge j incorrectly releases defendants of race r with misconduct potential (her "false negative rate"). Equation (1) can then be written:

$$\Delta_j = \left(\delta_{jw}^T - \delta_{jb}^T\right)(1 - \bar{\mu}) + \left(\delta_{jw}^F - \delta_{jb}^F\right)\bar{\mu} \tag{2}$$

where $\bar{\mu} = E[Y_i^*]$ denotes the overall risk of pretrial misconduct in the population of white and Black defendants. Equation (2) shows that Δ_j is a weighted average of racial differences in true and false negative rates for judge j. Since $1 - \delta_{jr}^T = Pr(D_{ij} = 0 \mid Y_i^* = 0, R_i = r)$ denotes the probability that judge j incorrectly detains defendants of race r without misconduct potential (her "false positive rate" for this race), Equation (2) also shows that Δ_j captures racial differences in type-I and type-II error rates. The system-wide level of discrimination similarly captures the case-weighted average racial difference in error rates across all judges.

The Δ_j parameters capture the differential treatment of Black and white defendants for all reasons unrelated to an individual's true potential for pretrial misconduct, a measure that is consistent with mainstream notions of discrimination in the legal, economic, and computer science literatures. The intentional unequal treatment of otherwise identical Black and white individuals is prohibited by the Equal Protection Clause of the 14th Amendment, and, more generally, is unwarranted because membership in a particular demographic group is not relevant to the purposes of the criminal justice system (Yang and Dobbie, 2019). Estimating Δ_j is therefore an important first step to establish unconstitutional behavior by judges in many cases, though it may not be sufficient absent proof of discriminatory intent.[8] Our measure also aligns with the labor market definition of discrimination

[8]The Supreme Court has ruled that "official action will not be held unconstitutional [under the Equal Protection Clause] solely because it results in a racially disproportionate impact...Proof of racially discriminatory intent or purpose

in Aigner and Cain (1977), which compares the treatment of white and Black workers with the same objective level of productivity. We analogously compare the release rates of white and Black defendants with the same objective potential for pretrial misconduct, Y_i^*.[9] Finally, our measure is closely related to the idea of "conditional procedure accuracy equality" in the literature on algorithmic discrimination (Berk et al., 2018). This fairness condition imposes the equality of type-I and type-II error rates across race, which per Equation (2) would imply $\Delta_j = 0$.

Three further comments on the interpretation of Δ_j are warranted. First, the Δ_j parameters capture a broad notion of discrimination arising from either racial bias or statistical discrimination, as we formalize below. Both forms of discrimination may arise either because of how a judge directly considers defendant race or because of how a judge considers observable characteristics that are correlated with race. Judges may, for example, be excessively strict when a defendant is charged with certain drug offenses (relative to their relevance for future misconduct) and Black defendants may be more likely to be charged with these types of crimes. Judges may similarly place excessive weight on whether a defendant lives or works in a predominantly Black neighborhood (again relative to its relevance for future misconduct). In both cases, the Δ_j parameters will capture de facto discrimination through these seemingly race-neutral characteristics by showing that Black defendants are more likely to be detained than white defendants of equal misconduct potential.

Second, and relatedly, Δ_j is intended to capture the differential treatment of white and Black defendants with the same unobserved misconduct potential, not those with the same observable characteristics. Consider an extreme version of neighborhood-based discrimination in which white and Black defendants are identical except that white defendants are idiosyncratically more likely to live in a particular neighborhood. Suppose, in behavior akin to the illegal practice of redlining, judge j releases a higher proportion of defendants in predominantly white neighborhoods but within each neighborhood she is "race-blind," releasing white and Black defendants at the same rate. Conditional on neighborhood, we would find no release rate disparity. Yet, there is clearly discrimination as white and Black defendants have identical misconduct risk but Black defendants are detained at higher rates (so $\Delta_j > 0$). This example shows how adjusting for observable characteristics (such as neighborhood) need not bring observed disparities closer to Δ_j, a point we return to in Section 3.3.

Finally, we note that Δ_j captures discrimination at a single stage of the criminal justice system, but it can be affected by discrimination both at other points in the criminal justice system and in society as a whole. For example, potential over-policing of Black neighborhoods relative to white neighborhoods may impact the types of crimes that are reported to and investigated by the police, subsequently impacting the types of cases that a judge hears. We hold this population of cases fixed in measuring discrimination in bail decisions, and therefore hold fixed any unwarranted disparities at other points in the system that might affect Δ_j.

is required." (Arlington Heights v. Metropolitan Housing Development Corp., 429 U.S. 252, 264-65, 1977). In McCleskey v. Kemp, for example, the Court rejected a challenge to Georgia's capital punishment system despite statistical evidence of racial disparities in death penalty decisions because the evidence was "clearly insufficient to support an inference that any of the decisionmakers...acted with discriminatory purpose." 481 U.S. 279, 281-82 (1987).

[9] By comparison, Phelps (1972) suggests measuring discrimination by comparing the treatment of white and Black workers with the same subjective signal of labor market productivity. Discrimination measures based on objective potential outcomes (as in this paper and Aigner and Cain (1977)) and subjective signals of potential outcomes (as in Phelps (1972)) are generally the same when the quality of the signals is identical by race, but can differ when individuals of different races tend to generate more or less informative signals. We return this issue in Section 6, where we estimate a structural model that allows for more or less informative risk signals for defendants of different races.

3.2 Theoretical Drivers of Discrimination

Racial discrimination in the sense of $\Delta_j \neq 0$ can arise from two distinct theoretical channels. The first is racial bias, in which judges discriminate against Black defendants at the margin of pretrial release due to either racial preferences (Becker, 1957) or some form of inaccurate racial stereotyping (Bordalo et al., 2016). The second is statistical discrimination, in which judges act on accurate risk predictions but discriminate due to racial differences in average risk or the precision of received risk signals (Phelps, 1972; Arrow, 1973; Aigner and Cain, 1977). Racial discrimination can be defined and measured without a model for judicial decision-making, but understanding these theoretical channels requires imposing more structure on the binary release decisions D_{ij}.

We formalize the relationship between racial discrimination, racial bias, and statistical discrimination by considering a population of white and Black defendants assigned to a single bail judge. Following the classic analysis of Aigner and Cain (1977), we suppose that the judge observes each defendant's race R_i and a noisy signal of pretrial misconduct $\nu_i = Y_i^* + \eta_i$ with normally distributed noise: $\eta_i \mid Y_i^*, (R_i = r) \sim N(0, \sigma_r^2)$. We allow both average misconduct risk $\mu_r = E[Y_i^* \mid R_i = r]$ and the "quality" (i.e. precision) of risk signals $\tau_r = 1/\sigma_r$ to vary by defendant race $r \in \{w, b\}$. We assume the judge forms a posterior risk prediction $p(\nu_i, R_i)$ from the signal and the defendant's race. Here we assume this prediction is accurate, in that $p(\nu_i, R_i) = Pr(Y_i^* = 1 \mid \nu_i, R_i)$. We also assume the judge has a subjective benefit of releasing individuals of race r, given by $\pi_r \in (0, 1)$. The judge then releases all defendants whose benefit exceeds the posterior risk cost, yielding the decision rule:

$$D_i = \mathbf{1}[\pi_{R_i} \geq p(\nu_i, R_i)] \tag{3}$$

Appendix B.2 derives the specific form of the posterior function $p(\cdot)$, completing the model.[10]

Racial bias in the sense of Becker (1957) arises when the judge perceives a lower benefit from releasing Black defendants relative to white defendants with the same risk posterior, so that $\pi_b < \pi_w$. All else equal, such bias will lead to racial discrimination. By applying different thresholds to posterior risk, the judge generally makes different decisions for white and Black defendants with the same misconduct potential Y_i^*. If, for example, $\pi_b < \pi_w$ but both mean risk μ_r and signal quality τ_r are the same across race (implying a common distribution of $p(\nu_i, R_i)$ given Y_i^*), the judge will release fewer Black defendants conditional on Y_i^* such that $\Delta_j > 0$. Inaccurate racial stereotyping can similarly result in discrimination and tends to be observationally equivalent to such racial animus (Arnold, Dobbie and Yang, 2018). In this case, even though judges believe they are applying the same threshold (i.e. $\pi_b = \pi_w$), inaccurate posterior beliefs lead them to effectively set different release standards by race. Since inaccurate stereotyping and racial animus tend to be observationally equivalent, we use the term racial bias for both potential channels. Inaccurate stereotyping and racial animus can both manifest through seemingly-race neutral characteristics, such as crime type or neighborhood. In the redlining example described above, for instance, a judge may indirectly set race-specific thresholds by prioritizing defendants from predominantly white neighborhoods.

Statistical discrimination in the sense of Aigner and Cain (1977) arises when judges act on accurate race-specific predictions of defendant risk but discriminate because these predictions are affected by racial differences in either the average misconduct risk μ_r or signal quality τ_r. Differences in average

[10] An alternative model of judicial decision-making specifies race-specific costs of misconduct classification errors. Appendix B.3 shows how such a model also leads to a threshold decision rule, with π_r denoting the relative cost of releasing defendants with misconduct potential.

misconduct risk μ_r will tend to lead to lower release rates for defendants in the group with higher average misconduct risk, thereby resulting in discrimination against that group. Suppose, for example, that signal quality and release benefits are the same across race ($\tau_b = \tau_w$ and $\pi_b = \pi_w$) but the average level of risk is higher for Black defendants ($\mu_b > \mu_w$). The judge uses both the risk signal ν_i and the defendant's race to accurately predict misconduct, so the judge's posterior $p(\nu_i, R_i)$ will be systematically higher among Black defendants given ν_i. Black defendants will thus be less likely to be released conditional on Y_i^*, such that $\Delta_j > 0$, even though the judge's posterior threshold π_r and the distribution of risk signals ν_i do not depend on race given Y_i^*. Statistical discrimination due to differences in signal quality τ_r will instead have an ambiguous effect on release rates disparities. If, for example, a judge's release threshold π_r is higher than the average level of misconduct risk in the population μ_r then noisier risk signals will lead to fewer defendants of that race being detained given true misconduct potential, as judges place more weight on the mean risk μ_r which falls below the threshold.[11] Differences in signal quality may reflect differences in the informativeness of seemingly race-neutral characteristics, such as when a defendant's neighborhood is more predictive of pretrial misconduct potential for white defendants relative to Black defendants.

Racial bias and accurate statistical discrimination can both generate unwarranted release rate disparities, but these two theoretical drivers yield different predictions for misconduct outcomes at the margin of release. When risk posteriors are accurate, marginal released outcomes capture the race-specific release benefits:

$$E[Y_i^* \mid p(\nu_i; R_i) = \pi_{R_i}, R_i] = E[Y_i^* \mid E[Y_i^* \mid \nu_i, R_i] = \pi_{R_i}, R_i] = \pi_{R_i} \tag{4}$$

Marginal white and marginal Black defendants should therefore have the same misconduct rate at the margin of release if the judge is racially unbiased ($\pi_w = \pi_b$), but marginal white defendants should have a higher probability of misconduct if the judge is racially biased against Black defendants ($\pi_w > \pi_b$). Finding unequal marginal outcomes thus rejects accurate statistical discrimination as the sole reason for finding $\Delta_j \neq 0$.

3.3 Empirical Challenges

Estimating racial discrimination is difficult because observational comparisons of white and Black release rates cannot control for unobserved misconduct potential and are therefore likely to suffer from omitted variables bias (OVB). Testing for specific drivers of discrimination, such as racial bias, is also difficult unless judges have a common ordering of defendants by their appropriateness for release, satisfying a conventional but strong assumption of first-stage monotonicity.

To formalize these empirical challenges, we introduce new notation for the data observed by an econometrician. Let $Z_{ij} = 1$ if defendant i is assigned to judge j, let $D_i = \sum_j Z_{ij} D_{ij}$ indicate the defendant's release status, and let $Y_i = D_i Y_i^*$ indicate the observed pretrial misconduct outcome. The expression for observed pretrial misconduct reflects the fact that an individual who is detained ($D_i = 0$) cannot fail to appear in court or be rearrested for a new crime, such that $Y_i = 0$ when $D_i = 0$ despite individual i's pretrial misconduct potential Y_i^*. The econometrician observes $\left(R_i, (Z_{ij})_{j=1}^J, D_i, Y_i\right)$ for

[11] The theoretical literature typically considers racial bias and statistical discrimination in isolation, while our empirical analysis allows racial differences in risk thresholds π_r, signal quality τ_r, and mean risk μ_r to each affect unwarranted disparity Δ_j. We continue to refer to the case of $\pi_w \neq \pi_b$ as racial bias in the model, while referring to $\tau_w \neq \tau_b$ or $\mu_w \neq \mu_b$ as statistical discrimination.

each defendant, and records whether the defendant is white in an indicator, $W_i = \mathbf{1}[R_i = w]$.

Rotational assignment of arraignment shifts can generate quasi-random assignment of individuals to different bail judges. To show how such quasi-experimental variation can and cannot help with measuring racial discrimination and its drivers we assume here that judges are simply randomly assigned, such that Z_{ij} is independent of (R_i, D_{ij}, Y_i^*) for each j. In practice, we relax this assumption to allow for the conditional quasi-random assignment in our setting.

Omitted Variables Bias in Observational Comparisons

Observational disparity analyses, whether in bail decisions or another area of the criminal justice system, usually come from "benchmarking" regressions of outcomes such as pretrial release on an indicator for an individual's race and controls for the observed characteristics of those individuals (e.g., Gelman, Fagan and Kiss, 2007; Abrams, Bertrand and Mullainathan, 2012). Since pretrial misconduct potential Y_i^* is both unobserved and likely to affect release decisions, such observational comparisons are likely to produce biased estimates of the discrimination parameters Δ_j.

To formalize the OVB challenge, we consider a simple judge-specific benchmarking regression:

$$D_i = \sum_j \alpha_j W_i Z_{ij} + \sum_j \phi_j Z_{ij} + \epsilon_i \tag{5}$$

where D_i is again an indicator for pretrial release, $W_i Z_{ij}$ is the interaction of the indicator for the defendant being white and a fixed effect of each judge, and Z_{ij} are non-interacted judge fixed effects. We omit the constant term so that all judge fixed effects are included, and abstract away from other defendant observables for simplicity. The interaction coefficients thus measure the difference in judge j's release rates for white defendants relative to Black defendants:

$$\alpha_j = E[D_i \mid R_i = w, Z_{ij} = 1] - E[D_i \mid R_i = b, Z_{ij} = 1] \tag{6}$$

While we focus here on a judge-specific benchmarking regression, the same conclusions emerge from an analysis of a simpler system-wide benchmarking regression of $D_i = \phi + \alpha W_i + \epsilon_i$.

Even with random judge assignment, the release rate disparities α_j will tend to differ from the unwarranted disparity parameters Δ_j. When Z_{ij} is independent of (R_i, D_{ij}, Y_i^*),

$$\alpha_j = E[D_{ij} \mid R_i = w] - E[D_{ij} \mid R_i = b] \tag{7}$$

Defining, as above, $\mu_r = E[Y_i^* \mid R_i = r]$ as the average misconduct risk among individuals of race r and $(\delta_{jr}^T, \delta_{jr}^F)$ as the judge's true and false negative rates for individuals of race r, these release rate disparities can be written:

$$\alpha_j = \left(\delta_{jw}^T(1-\mu_w) + \delta_{jw}^F \mu_w\right) - \left(\delta_{jb}^T(1-\mu_b) + \delta_{jb}^F \mu_b\right) \tag{8}$$

In contrast, judge j's unwarranted release rate disparity given by Equation (2) can be written:

$$\Delta_j = \left(\delta_{jw}^T(1-\bar{\mu}) + \delta_{jw}^F \bar{\mu}\right) - \left(\delta_{jb}^T(1-\bar{\mu}) + \delta_{jb}^F \bar{\mu}\right) \tag{9}$$

where $\bar{\mu} = E[Y_i^*] = p_w \mu_w + p_b \mu_b$ is the average misconduct risk in the population of defendants, with

$p_r = Pr(R_i = r)$ denoting racial shares.

The difference between the benchmarking regression coefficient α_j in Equation (8) and the judge discrimination measure Δ_j in Equation (9) measures OVB in the simple benchmarking regression given by Equation (5). This difference can be written

$$\begin{aligned}\xi_j \equiv \alpha_j - \Delta_j &= \left(\delta^T_{jw}(\bar{\mu} - \mu_w) + \delta^F_{jw}(\mu_w - \bar{\mu})\right) - \left(\delta^T_{jb}(\bar{\mu} - \mu_b) + \delta^F_{jb}(\mu_b - \bar{\mu})\right) \\ &= (\mu_b - \mu_w) \times \left[\left(\delta^T_{jw} - \delta^F_{jw}\right) p_b + \left(\delta^T_{jb} - \delta^F_{jb}\right) p_w\right] \end{aligned} \tag{10}$$

where the second line follows by definition of the population risk $\bar{\mu}$. The regression coefficient α_j will be biased upward for Δ_j when $\xi_j > 0$ and biased downward when $\xi_j < 0$.

Three insights follow from the OVB formula (10). First, conventional benchmarking regressions will generally yield biased estimates of the absolute level of discrimination Δ_j, even with quasi-random judge assignment. The exception is when either judge release decisions are independent of potential misconduct for each race (i.e., $E[D_{ij} \mid Y_i^*, R_i]$ does not depend on Y_i^*, so $\delta^T_{jr} = \delta^F_{jr}$) or mean misconduct risk is identical across race (i.e., $\mu_w = \mu_b$). Both scenarios are unlikely in practice.

Second, conventional benchmarking regressions will also yield biased estimates of the relative differences in the extent of racial discrimination across judges, even when judges are as-good-as-randomly assigned. The extent of OVB can also vary across judges in Equation (10), such that difference in benchmarking coefficients between judge j and k identifies $\alpha_j - \alpha_k = \Delta_j - \Delta_k + \xi_j - \xi_k$ and not $\Delta_j - \Delta_k$. In general, OVB will vary across judges whenever there are differential responses to misconduct potential differences, such that $\delta^T_{jr} - \delta^F_{jr}$ varies across j for either race r.[12]

Third, Equation (10) suggests a potential avenue for estimating racial discrimination when bail judges are as-good-as-randomly assigned, using familiar econometric objects. One of the terms driving the OVB of each α_j is the difference in race-specific misconduct risk in the population, $\mu_b - \mu_w$, which is common to all judges. With Y_i^* capturing defendant i's potential for pretrial misconduct when released and $Y_i = 0$ for all detained individuals, the $\mu_r = E[Y_i^* \mid R_i = r]$ can be understood as average treatment effects (ATEs), of pretrial release on pretrial misconduct, among individuals of race r. We show in Section 5 how such ATEs can be estimated from quasi-experimental judge assignment and used to purge OVB from conventional benchmarking estimates, recovering valid estimates of Δ_j.

It is also worth highlighting that adding observable characteristics to the simple benchmarking regression (5) need not solve the OVB challenge and may either bring observed disparities closer to or further away from Δ_j (Ayres, 2010). Even when the observables are rich enough to absorb differences in misconduct potential Y_i^*, their inclusion may introduce new bias by absorbing part of the judge's decision-making process that yields discrimination on seemingly race-neutral characteristics (as in our redlining example above). This observation cautions against an approach that conditions on as many observables as possible in order to "explain away" observed racial disparities, or which interprets a covariate-adjusted disparity as an upper bound on racial discrimination.

[12] To see this simply, suppose all judges are non-discriminatory, with $\delta^T_{jr} = \delta^T_j$ and $\delta^F_{jr} = \delta^F_j$ for each j and r, such that $\Delta_j = 0$ for each j. Suppose further that judges release all defendants without misconduct potential, such that $\delta^T_j = 1$. Differences in judge leniency are then solely due to differences in their rate of releasing defendants with misconduct potential, δ^F_j. Equation (10) shows that these differences drive differences in OVB, since $\xi_j = (\mu_b - \mu_w)(1 - \delta^F_j)$ in this case. Consequently, a benchmarking analysis would tend to incorrectly suggest not only racial discrimination ($\xi_j > 0$) but also differential discrimination across judges ($\xi_j \neq \xi_k$) when the average risk differs by race.

Monotonicity Violations in Standard IV Estimates

Testing for racial bias and statistical discrimination is also difficult unless judges have a common ordering of defendants by their appropriateness for release, satisfying a conventional first-stage monotonicity assumption. For example, standard IV methods can be used to test for racial bias (whether due to preferences or inaccurate stereotyping) given the quasi-random assignment of judges and first-stage monotonicity (Arnold, Dobbie and Yang, 2018; Marx, 2018). Monotonicity is, however, an especially strong assumption in this setting, implying that all judges are equally skilled in predicting an individual's propensity for pretrial misconduct and only differ in terms of the thresholds they set on a common posterior risk ordering.

To illustrate this potential limitation of the standard IV-based test for racial bias, we consider a multiple-judge generalization of the earlier decision model. The release rule for each judge j is given by $D_{ij} = \mathbf{1}[\pi_{jR_i} \geq p_j(\nu_{ij}, R_i)]$, where π_{jr} is the race-specific release benefit of judge j, and $p_j(v, r)$ is the judge's posterior for the misconduct risk of a defendant of race r who sends a signal of v. The most general version of this model allows risk posteriors to differ across judges because of heterogeneous beliefs and signal qualities. Correspondingly, we index the signals ν_{ij} of heterogenous quality τ_{jr} by j as well as by r. Judges with higher τ_{jr} can be thought of as being more skilled, in that they base decisions on more predictive signals of misconduct potential.

Conventional first-stage monotonicity identifies marginal misconduct outcomes for white and Black defendants, which can be used to test for racial bias by assuming judges form common risk posteriors. Per Imbens and Angrist (1994), when $p_j(\nu_{ij}, R_i) = p(\nu_i, R_i)$ does not vary by j, a linear IV regression of misconduct outcomes Y_i on release D_i instrumented by quasi-randomly assigned judge indicators Z_{ij}, in a sample of either white or Black individuals assigned to one of two judges, identifies a local average treatment effect:

$$\frac{E[Y_i \mid Z_{ij} = 1, R_i] - E[Y_i \mid Z_{ik} = 1, R_i]}{E[D_i \mid Z_{ij} = 1, R_i] - E[D_i \mid Z_{ik} = 1, R_i]} = E[Y_i^* \mid \pi_{jR_i} \geq p(\nu_i, R_i) > \pi_{kR_i}] \tag{11}$$

where here the effect of "treating" individual i with release is simply her misconduct potential Y_i^*. Equation (11) thus gives the average misconduct risk for "compliers" of race R_i, whose posterior risk $p(\nu_i, R_i)$ lies between the two judge benefit thresholds π_{jR_i} and π_{kR_i} (where $\pi_{jR_i} \geq \pi_{kR_i}$ without loss). As these two thresholds become closer, the IV estimand in Equation (11) approaches the marginal released outcomes of each judge in Equation (4) and can therefore be used to test whether $\pi_{jw} = \pi_{jb}$. Arnold, Dobbie and Yang (2018) show how standard linear and local IV procedures yield such tests in settings with many quasi-randomly assigned bail judges.

When judge skill varies, however, first-stage monotonicity is violated and standard IV procedures may not capture average misconduct risk for marginal defendants. If $\tau_{jr} \neq \tau_{kr}$, then $p_j(\nu_{ij}, R_i) \neq p_k(\nu_{ik}, R_i)$ and the same linear IV regression instead identifies a non-convex linear combination of treatment effects for "complier" and "defier" populations:

$$\begin{aligned}\frac{E[Y_i \mid Z_{ij} = 1, R_i] - E[Y_i \mid Z_{ik} = 1, R_i]}{E[D_i \mid Z_{ij} = 1, R_i] - E[D_i \mid Z_{ik} = 1, R_i]} &= p_{cR_i} E[Y_i^* \mid \pi_{jR_i} \geq p_j(\nu_{ij}, R_i), p_k(\nu_{ik}, R_i) > \pi_{kR_i}] \\ &\quad - p_{dR_i} E[Y_i^* \mid \pi_{kR_i} \geq p_k(\nu_{ik}, R_i), p_j(\nu_{ij}, R_i) > \pi_{jR_i}]\end{aligned} \tag{12}$$

where p_{cr} is proportional to the complier share of the population of race r who is newly released when

switching assignment from judge k to judge j, and p_{dr} is proportional to the defier share who is newly detained (with $p_{cr} - p_{dr} = 1$). Unlike Equation (11), Equation (12) generally cannot be used to isolate marginal released outcomes and test whether $\pi_{jw} = \pi_{jb}$. Consequently, the IV-based tests for racial bias proposed by Arnold, Dobbie and Yang (2018) are generally invalid when judge skill varies. In Section 6, we develop an alternative approach to test for racial bias in a model that allows for variation in judge skill. We also show how statistical discrimination due to racial differences in average risk or signal quality across race can be measured in this more realistic model with heterogeneous judge skill.

4 Data and Observational Comparisons

4.1 Sample and Summary Statistics

Our analysis of racial discrimination in bail decisions is based on the universe of 1,458,056 arraignments made in NYC between November 1, 2008 and November 1, 2013. The data contain information on a defendant's gender, race, date of birth, and county of arrest, as well as the (anonymized) identity of the assigned bail judge. In our primary analysis, we categorize defendants as white (including both non-Hispanic and Hispanic white individuals), Black (including both non-Hispanic and Hispanic Black individuals), or neither. We explore alternative categorizations of race in robustness checks below.

In addition to detailed demographics, our data contain information on each defendant's current offense, history of prior criminal convictions, and history of past pretrial misconduct (both rearrests and FTA). We also observe whether the defendant was released at the time of arraignment and whether this release was due to release without conditions or some form of money bail. We categorize defendants as either released (including both release without conditions and with paid cash bail) or detained (including cash bail that is not paid) at the first arraignment, though we again explore robustness to other categorizations of the initial pretrial release decision below. Finally, we observe whether a defendant subsequently failed to appear for a required court appearance or was subsequently arrested for a new crime before case disposition. We take either form of pretrial misconduct as the primary outcome of our analysis, but again explore robustness to other measures below.

We make four key restrictions to arrive at our estimation sample. First, we drop cases where the defendant is not charged with a felony or misdemeanor (N=26,057). Second, we drop cases that were disposed at arraignment (N=364,051) or adjourned in contemplation of dismissal (N=230,517). This set of restrictions drops cases that are likely to be dismissed by virtually every judge: Appendix Table A1 confirms that judge assignment is not systematically related to case disposal or case dismissal. Third, we drop cases in which the defendant is assigned a cash bail of \$1 (N=1,284). This assignment occurs in cases in which the defendant is already serving time in jail on an unrelated charge; the \$1 cash bail is set so that the defendant receives credit for served time, and does not reflect a new judge decision. Fourth, we drop defendants who are non-white and non-Black (N=45,529). Finally, we drop defendants assigned to judges with fewer than 100 cases (N=3,785) and court-by-time cells with fewer than 100 cases, only one unique judge, or only Black or only white defendants for a given judge (N=191,647), where a court-by-time cell is defined by the assigned courtroom, shift, day-of-week, month and year (e.g., the Wednesday night shift in Courtroom A of the Kings County courthouse in January 2012). The final sample consists of 595,186 cases, 367,434 defendants, and 268 judges.[13]

[13] Appendix Table A2 compares the full sample of NYC bail cases to our estimation sample. By construction, our estimation sample has a somewhat lower release rate, although the ratio of release rates by race is similar. Our estimation

Table 1 summarizes our estimation sample, both overall and by race. Panel A shows that 73.0 percent of defendants are released before trial. A defendant is defined as released before trial if either the defendant is released without conditions (ROR) or the defendant posts the required bail amount before disposition. The vast majority of these releases are without conditions, with only 14.4 percent of defendants being released after being assigned money bail. White defendants are more likely to be released before trial than Black defendants, with a 76.7 percent release rate relative to a 69.5 percent release rate. Among released defendants, however, the distribution of release conditions (e.g. the ROR share) is virtually identical across race.

Judges may release white defendants at a higher rate than Black defendants because of relevant differences in defendant or charge characteristics. Consistent with this idea, Panel B of Table 1 shows that Black defendants are 4.9 percentage points more likely to have been arrested for a new crime before trial in the past year compared to white defendants, as well as 3.0 percentage points more likely to have a prior FTA in the past year. Panel C further shows that Black defendants are 1.3 percentage points more likely to have been charged with a felony compared to white defendants, as well as 3.6 percentage points more likely to have been charged with a violent crime. Finally, Panel D shows that Black defendants who are released are 6.6 percentage points more likely to be rearrested or have an FTA than white defendants who are released (though the composition of such misconduct is similar). Importantly, and in contrast to the other statistics in Table 1, the risk statistics in Panel D are only measured among released defendants. Pretrial misconduct potential is, by definition, unobserved among detained individuals despite being the key legal objective for bail judges.

4.2 Quasi-Experimental Judge Assignment

Our empirical strategy exploits variation in pretrial release from the quasi-random assignment of judges who vary in the leniency of their bail decisions. There are three features of the NYC pretrial system that make it an appropriate setting for this research design.

First, NYC uses a rotation calendar system to assign judges to arraignment shifts in each of the five county courthouses in the city, generating quasi-random variation in bail judge assignment for defendants arrested at the same time and in the same place. Each county courthouse employs a supervising judge to determine the schedule that assigns bail judges to the day (9 a.m. to 5 p.m.) and night arraignment shift (5 p.m. to 1 a.m.) in one or more courtrooms within each courthouse. Individual judges can request to work certain days or shifts but, in practice, there is considerable variation in judge assignments within a given arraignment shift, day-of-week, month, and year cell.

Second, there is limited scope for influencing which bail judge will hear any given case, as most individuals are brought for arraignment shortly after their arrest. Each defendant's arraignment is also scheduled by a coordinator, who seeks to evenly distribute the workload to each open courtroom at an arraignment shift. Combined with the rotating calendar system described above and the processing time required before the arraignment, it is unlikely that police officers, prosecutors, defense attorneys, or defendants could accurately predict which judge is presiding over any given arraignment.

Finally, the rotation schedule used to assign bail judges to cases does not align with the schedule of any other actors in the criminal justice system. For example, different prosecutors and public defenders handle matters at each stage of criminal proceedings and are not assigned to particular

sample is also broadly representative in terms of defendant and charge characteristics, with a slightly higher share of defendants with prior FTAs and rearrests, and a lower share of defendants charged with drug and property crimes.

bail judges, while both trial and sentencing judges are assigned to cases via different processes. As a result, we can study the effects of being assigned to a given bail judge as opposed to, for example, the effects of being assigned to a given set of bail, trial, and sentencing judges.

Appendix Table A3 verifies the quasi-random assignment of judges to bail cases in the estimation sample. Each column reports coefficient estimates from an ordinary least squares (OLS) regression of judge leniency on various defendant and case characteristics, with court-by-time fixed effects that control for the level of quasi-experimental bail judge assignment. We measure leniency using the leave-one-out average release rate among all other defendants assigned to a defendant's judge, following the standard approach in the literature (e.g., Arnold, Dobbie and Yang, 2018; Dobbie, Goldin and Yang, 2018). Most coefficients in this balance table are small and not statistically significantly different from zero, both overall and by defendant race. A joint F-test fails to reject the null of quasi-random assignment at conventional levels of statistical significance.[14]

Appendix Table A4 further verifies that the assignment of different judges meaningfully affects the probability an individual is released before trial. Each column of this table reports coefficient estimates from an OLS regression of an indicator for pretrial release on judge leniency and court-by-time fixed effects. A one percentage point increase in the predicted leniency of an individual's judge leads to a 0.96 percentage point increase in the probability of release, with a somewhat smaller first-stage effect for white defendants and a somewhat larger effect for Black defendants.

4.3 Observational Comparisons

Table 2 investigates the system-wide level of observed racial disparity in NYC pretrial release rates. We estimate OLS regressions of the form:

$$D_i = \phi + \alpha W_i + X_i'\beta + \epsilon_i \tag{13}$$

where D_i is an indicator equal to one if defendant i is released, W_i is an indicator for the defendant being white, and X_i is a vector of controls. Column 1 of Table 2 omits any controls in X_i, column 2 adds court-by-time fixed effects to adjust for unobservable differences at the level of quasi-experimental bail judge assignment to X_i, and column 3 further adds the defendant and case observables from Table 1. Such regressions generally follow the conventional benchmarking approach from the literature (e.g., Gelman, Fagan and Kiss, 2007; Abrams, Bertrand and Mullainathan, 2012); we again note that the defendant and case observables included in column 3 can either increase or decrease OVB.

Table 2 documents both statistically and economically significant release rate disparities between white and Black defendants in NYC. The unadjusted white-Black release rate difference α is estimated in column 1 at 7.2 percentage points, with a standard error (SE) of 0.5 percentage points. This release rate gap is around 10 percent of the mean release rate of 73 percent. The release rate gap falls slightly, to 6.8 percentage points (SE: 0.5), when we control for court-by-time fixed effects. The gap falls by an additional 24 percent, to 5.2 percentage points (SE: 0.4), when we add defendant and

[14] Even with the quasi-random assignment of bail judges, the exclusion restriction in our framework could be violated if judge assignment impacts the probability of pretrial misconduct through channels other than pretrial release. While the assumption that judges only systematically affect defendant outcomes through pretrial release is fundamentally untestable, we join Arnold, Dobbie and Yang (2018) in viewing it as reasonable here. Bail judges only handle one decision, limiting the potential channels through which they could affect defendants. Pretrial misconduct is also a relatively short-run outcome, further limiting the role of alternative channels. In a similar setting, Dobbie, Goldin and Yang (2018) find that there are no independent effects of the assigned money bail amount on defendant outcomes. We explore the robustness of our findings to such effects below.

case observables. These estimates are similar in magnitude to the association, reported in column 3, between the probability of release and having an additional drug charge (-5.7 percentage points) or pretrial arrest (-6.8 percentage points) in the past year.

Figure 1 summarizes the distribution of judge-specific release rate disparities across the 268 bail judges in our sample. We estimate judge-specific disparities from OLS regressions of the form:

$$D_i = \sum_j \alpha_j W_i Z_{ij} + \sum_j \phi_j Z_{ij} + X_i'\beta + \epsilon_i \tag{14}$$

where D_i is again an indicator equal to one if defendant i is released, $W_i Z_{ij}$ is the interaction between an indicator for the defendant being white and the fixed effects for each judge, Z_{ij} are the non-interacted fixed effects for each judge, and X_i is again a control vector. We first estimate Equation (14) with X_i demeaned, such that the α_j captures regression-adjusted difference in release rates for white and Black individuals assigned to judge j. We then compute empirical Bayes posteriors of α_j using standard shrinkage procedures (Morris, 1983). Figure 1 shows the distribution of the racial disparity posteriors that adjust only for the main judge fixed effects and court-by-time fixed effects, following column 2 of Table 2, as well as the distribution of posteriors when we add both defendant and case observables and court-by-time fixed effects, following column 3 of Table 2. Figure 1 also reports an estimate of the prior mean and standard deviation of α_j across judges, as well as the fraction of judges with positive α_j, via the posterior average effect approach of Bonhomme and Weidner (2020).[15]

The distributions of release rate disparity posteriors in Figure 1 are located well above zero, suggesting that nearly all judges in our sample release white defendants at a higher rate than Black defendants. We estimate that only 4.1 percent (SE: 1.3) of judges in our sample release a larger share of Black defendants in the specification that adjusts for court-by-time fixed effects, while only 5.9 percent (SE: 1.5) are estimated to release a larger share when we additionally adjust for defendant and case observables. Figure 1 nevertheless shows considerable variation in the magnitude of the release rate disparities across judges. The standard deviation of α_j is estimated at 4.0 percentage points (SE: 0.3) when we adjust for court-by-time fixed effects, and 3.3 percentage points (SE: 0.3) when we additionally adjust for defendant and case observables. The average judge-specific disparities, which differ from the system-wide averages in Table 2 due to differences in weighting, are 6.6 percentage points (SE: 0.2) when we adjust for court-by-time fixed effects, and 5.0 percentage points (SE: 0.2) when we additionally adjust for defendant and case observables.

The results from Table 2 and Figure 1 confirm large and pervasive racial disparities in NYC bail decisions, both in the raw data and after accounting for observable differences between white and Black defendants. These observational estimates suggest bail judges may be discriminating against Black defendants, but are not conclusive as we cannot directly adjust for unobserved misconduct potential Y_i^* and could thus either over- or understate the true level and distribution of discrimination in the NYC pretrial system. We next develop and apply a quasi-experimental approach to adjust for unobserved misconduct potential Y_i^* and remove OVB from these observational comparisons.

[15] See Appendix B.4 for the details of the conventional empirical Bayes procedures we apply in this section.

5 Quasi-Experimental Estimates of Racial Discrimination

5.1 Methods

We estimate racial discrimination in pretrial release decisions by rescaling the observational release rate comparisons in Figure 1 using quasi-experimental estimates of average white and Black misconduct risk. This quasi-experimental approach leverages first-stage variation in judge leniency but, unlike standard IV methods, does not require a first-stage monotonicity assumption. We only require that average misconduct risk among white and Black defendants can be extrapolated from the quasi-experimental data, and that the judges' legal objective is well-specified by the econometrician.

The first key insight underlying our approach is that when judges are as-good-as-randomly assigned, the problem of measuring unwarranted release rate disparities for individual judges is equivalent to the problem of estimating the average misconduct risk among the full population of Black and white defendants. The source of OVB in an observational benchmarking comparison is the correlation between race and unobserved misconduct potential among a given judge's pool of white and Black defendants. With quasi-random judge assignment, this correlation is common to all judges and captured by race-specific population misconduct risk. Thus, given estimates of these race-specific risk parameters, observed release outcomes can be appropriately rescaled to make released white and Black defendants comparable in terms of their unobserved misconduct potential.

The rescaling that purges OVB from observational comparisons is given by expanding the true and false negative rates from our definition of racial discrimination in Equation (2):

$$\delta^T_{jr} = E[D_{ij} \mid Y^*_i = 0, R_i = r] = \frac{E[D_{ij}(1 - Y^*_i) \mid R_i = r]}{E[1 - Y^*_i \mid R_i = r]} = \frac{E[D_i(1 - Y_i) \mid R_i = r, Z_{ij} = 1]}{1 - \mu_r} \tag{15}$$

$$\delta^F_{jr} = E[D_{ij} \mid Y^*_i = 1, R_i = r] = \frac{E[D_{ij}Y^*_i \mid R_i = r]}{E[Y^*_i \mid R_i = r]} = \frac{E[D_iY_i \mid R_i = r, Z_{ij} = 1]}{\mu_r} \tag{16}$$

where the third equalities in both lines follow from quasi-random judge assignment and the definition of mean risk $\mu_r = E[Y^*_i \mid R_i = r]$. Substituting these expanded true and false negative rates into Equation (2) yields:

$$\begin{aligned}\Delta_j =& E[D_i(1 - Y_i) \mid R_i = w, Z_{ij} = 1]\frac{1 - \bar{\mu}}{1 - \mu_w} + E[D_iY_i \mid R_i = w, Z_{ij} = 1]\frac{\bar{\mu}}{\mu_w} \\ &- E[D_i(1 - Y_i) \mid R_i = b, Z_{ij} = 1]\frac{1 - \bar{\mu}}{1 - \mu_b} - E[D_iY_i \mid R_i = b, Z_{ij} = 1]\frac{\bar{\mu}}{\mu_b} \\ =& E[\Omega_iD_i \mid R_i = w, Z_{ij} = 1] - E[\Omega_iD_i \mid R_i = b, Z_{ij} = 1]\end{aligned} \tag{17}$$

where:

$$\Omega_i = (1 - Y_i)\frac{1 - \bar{\mu}}{1 - \mu_{R_i}} + Y_i\frac{\bar{\mu}}{\mu_{R_i}} > 0 \tag{18}$$

The rewritten definition of discrimination in Equation (17) shows that judge j's level of discrimination Δ_j is given by the α_j coefficients in a simple benchmarking regression, where the release decisions D_i of each individual are rescaled by a positive factor Ω_i. This Ω_i reweights the sample to make released white and Black defendants comparable in terms of their unobserved misconduct potential. It therefore reveals the extent to which each judge discriminates against white and Black defendants with

identical misconduct potential, even though misconduct potential is unobserved and cannot be directly conditioned on. Equation (18) shows that Ω_i is a function of observed misconduct outcomes Y_i and the unobserved average race-specific misconduct risk parameters μ_r, where again $\bar{\mu} = \mu_w p_w + \mu_b p_b$. The key econometric challenge is therefore to estimate average misconduct risk μ_r among the full population of white and Black defendants.

Appendix Table A5 uses a simple numerical example to illustrate how our rescaling approach allows us to measure discrimination in bail decisions, even though misconduct potential is unobserved and cannot be directly conditioned on. We suppose that there are two types of hypothetical defendants in our example, high-risk H types and low-risk L types. We assume the judge is type-neutral when making release decisions. If the defendant has $Y_i^* = 1$, then there is an 80 percent chance the defendant is released regardless of type. If the defendant has $Y_i^* = 0$, then there is a 20 percent chance the defendant is released regardless of her type. Thus, while the judge receives a signal of the defendant's unobserved misconduct potential, this signal is not perfectly predictive, implying the judge will release some defendants who will misbehave and detain some defendants that would not misbehave. We also assume that 75 of the 100 hypothetical H-type defendants have misconduct potential ($Y_i^* = 1$) but only 25 of the 100 hypothetical L-type defendants have misconduct potential, such that $\mu_H = 0.75$ and $\mu_L = 0.25$. Panel A shows that this judge therefore has a release rate of 0.65 for L-type defendants but a release rate of 0.35 for H-type defendants, meaning that a conventional benchmarking regression would find that L-type defendants have a 30 percentage point higher release rate than H-type defendants ($\alpha_j = 0.3$) despite the judge being type-neutral.

Panel B of Appendix Table A5 shows how discrimination can be measured in this simple numerical example with observational release rate comparisons that are rescaled using average misconduct risk. Following Equations (17) and (18), we compute $\Omega_i = \frac{0.50}{0.75} = 2/3$ for released H-type defendants with $Y_i = 0$ and released L-type defendants with $Y_i = 1$, and $\Omega_i = \frac{0.50}{0.25} = 2$ for released L-type defendants with $Y_i = 1$ and released H-type defendants with $Y_i = 0$. The rescaling factor thus up-weights the release rates of individuals who are relatively less common in each type (risky L-type defendants and non-risky H-type defendants), while down-weighting the release rates of individuals who are relatively more common (non-risky L-type defendants and risky H-type defendants).[16] In this way, the rescaling factor equalizes the proportion of risky and non-risky defendants by type, meaning that a rescaled benchmarking regression would correctly find that H- and L-type defendants with the same misconduct potential have identical release rates ($\Delta_j = 0$).

The second key insight underlying our approach is that the average race-specific misconduct risk parameters that enter Equation (17) can be estimated from quasi-experimental variation in pretrial release and misconduct rates. To build intuition for our approach, consider a setting with as-good-as-random judge assignment and a supremely lenient bail judge j^* who releases nearly all defendants regardless of their race or potential for pretrial misconduct. This supremely lenient judge's race-specific

[16] This pattern of up- and down-weighting generally arises when H-type defendants have higher misconduct risk: i.e., when $\mu_H > \bar{\mu} > \mu_L$. In such cases, observations of released L-type defendants who subsequently offend are up-weighted ($Y_i - \mu_L > 0$ and $\bar{\mu} - \mu_L > 0$ so $\Omega_i > 1$), as are observations of released H-type defendants who do not subsequently offend ($Y_i - \mu_H < 0$ and $\bar{\mu} - \mu_H < 0$, so again $\Omega_i > 1$). Equation (17) also shows that $\Delta_j = \alpha_j - (E[(1-\Omega_i)D_i \mid R_i = w, Z_{ij} = 1] - E[(1-\Omega_i)D_i \mid R_i = b, Z_{ij} = 1])$, so that our rescaling can be understood as subtracting OVB from the observational comparisons with OVB given by a $(1-\Omega_i)$-scaled release rate disparity. Conventional and rescaled benchmarking regressions are identical when average misconduct risk does not vary by type: if $\mu_H = \mu_L = \bar{\mu}$ then $\Omega_i = 1$ for all defendants.

release rate among both Black and white defendants is close to one:

$$E[D_i \mid Z_{ij^*} = 1, R_i = r] = E[D_{ij^*} \mid R_i = r] \approx 1 \tag{19}$$

making the race-specific misconduct rate among defendants she releases close to the race-specific average misconduct risk in the full population:

$$E[Y_i \mid D_i = 1, Z_{ij} = 1, R_i = r] = E[Y_i^* \mid D_{ij^*} = 1, R_i = r] \approx E[Y_i^* \mid R_i = r] = \mu_r \tag{20}$$

where the first equality in both expressions follows by quasi-random assignment. Without further assumptions, the decisions of a supremely lenient and quasi-randomly assigned judge can therefore be used to estimate the average misconduct risk parameters needed for our discrimination measure.

In the absence of such a supremely lenient judge, the required average misconduct risk parameters can be estimated using model-based or non-parametric extrapolations of release and misconduct rate variation across quasi-randomly assigned judges. This approach is analogous to the standard regression discontinuity approach of extrapolating average potential outcomes to a treatment cutoff from nearby observations. Here, released misconduct rates are extrapolated from quasi-randomly assigned judges to the release rate cutoff (of one) given by a hypothetical supremely lenient judge. Mean risk estimates may, for example, come from the vertical intercept, at one, of linear, quadratic, or local linear regressions of estimated released misconduct rates $E[Y_i^* \mid D_{ij} = 1, R_i = r]$ on estimated release rates $E[D_{ij} \mid R_i = r]$ across judges j within each race r. As we show below, extrapolations may also come from a model of judge behavior. Absent any extrapolations, conservative bounds on mean risk may be obtained from the released misconduct rates of highly (but not supremely) lenient judges. Each of these approaches build on recent advances in ATE estimation with multiple discrete instruments (e.g., Brinch, Mogstad and Wiswall, 2017; Mogstad, Santos and Torgovitsky, 2018; Hull, 2020) and a long literature on "identification at infinity" in sample selection models (e.g., Chamberlain, 1986; Heckman, 1990; Andrews and Schafgans, 1998). Importantly, they can be justified without a conventional monotonicity assumption, in contrast to some of the recent literature.[17]

5.2 Results

Mean Risk by Race

Figure 2 illustrates the quasi-experimental variation in judge release rates and released misconduct rates in NYC. The horizontal axis plots estimates of release rates $E[D_{ij} \mid R_i = r]$ for each judge j and each race r, obtained from the earlier benchmarking regression in Equation (14) that adjusts for court-by-time fixed effects. The vertical axis plots the corresponding estimates of released misconduct

[17] To see why a conventional monotonicity assumption is not needed to estimate mean risk by extrapolation, consider a simple model in which each judge's release decisions are given by $D_{ij} = \mathbf{1}[\kappa_j \geq \upsilon_{ij}]$ where $\upsilon_{ij} \mid \kappa_j, \lambda_j \sim U(0,1)$ without loss and $E[Y_i^* \mid \upsilon_{ij}, \kappa_j, \lambda_j] = \mu + \lambda_j(\upsilon_{ij} - \frac{1}{2})$. This model violates conventional monotonicity, since judges differ both in their orderings of individuals by the appropriateness of release (υ_{ij}) and their relative skill at predicting misconduct outcomes (λ_j). Nevertheless, when $E[\lambda_j \mid \kappa_j]$ is constant (linear) in κ_j, average released misconduct rates $E[Y_i^* \mid D_{ij} = 1, \kappa_j] = E[\mu + \frac{1}{2}\lambda_j(\kappa_j - 1) \mid \kappa_j]$ are linear (quadratic) in release rates $E[D_{ij}] = \kappa_j$, so that these extrapolations identify the ATE μ. More flexible extrapolations generally accommodate a broader range of judge decision-making models by leveraging richer quasi-experimental variation. In the limit, local linear regressions can yield non-parametric estimates of mean misconduct risk provided there are many lenient judges (Hull, 2020).

rates $E[Y_i^* \mid D_{ij} = 1, R_i = r]$, obtained from the analogous OLS regression:

$$Y_i = \sum_j \rho_j W_i Z_{ij} + \sum_j \zeta_j Z_{ij} + X_i'\gamma + u_i \tag{21}$$

estimated among released individuals ($D_i = 1$), where again X_i contains court-by-time fixed effects and is demeaned to include all judge indicators. These specifications leverage an auxiliary assumption of linear conditional expectations of D_{ij} and Y_i^* in order to tractably accommodate the conditional random assignment of bail judges given the court-by-time fixed effects.[18] We do not include other defendant and case observables in X_i, which will allow our subsequent estimates to capture discrimination on seemingly race-neutral characteristics. The necessary inclusion of court-by-time fixed effects implies, however, that we will be unable to detect indirect discrimination across courts or times of the day. For example, we would be unable to detect if judges tend to be stricter for all defendants in courts or times of that day that tend to see a higher share of Black defendants.

Figure 2 shows significant variation in race-specific release rates across judges, with several judges releasing a high fraction of their defendants for each race. Released misconduct rates tend to increase with judge leniency for both races, as would be predicted by a behavioral model in which the more lenient judges release riskier defendants at the margin. This pattern is shown by the two solid lines in Figure 2, representing the race-specific lines-of-best-fit through the quasi-experimental data. The lines-of-best-fit are obtained by OLS regressions of judge-specific released misconduct rate estimates on judge-specific release rate estimates, with the judge-level regressions weighted inversely by the variance of misconduct rate estimation error. We also plot curves-of-best-fit from judge-level quadratic and local linear specifications as dashed and dotted lines, respectively, with both specifications again weighted inversely by the variance of misconduct rate estimation error. The simple linear specification fits the local IV variation well, with quadratic and local linear specifications yielding similar fits across much of the leniency distribution.

The vertical intercepts of the different curves-of-best-fit, at one, provide different estimates of the race-specific mean risk parameters μ_r. These estimates are reported in Panel A of Table 3. The simplest linear extrapolation, summarized in column 1, yields precise mean risk estimates of 0.338 (SE: 0.007) for white defendants and 0.400 (SE: 0.006) for Black defendants.[19] This extrapolation suggests that the average misconduct risk within the population of potential Black defendants is 6.2 percentage points higher than among the population of potential white defendants in this setting. Per the discussion in Section 3.3, such a racial gap in misconduct risk is likely to generate OVB in observational release rate comparisons.

The quadratic and local linear extrapolations of quasi-experimental variation yield similar race-specific mean risk estimates, as can be seen from Figure 2. The quadratic fit suggests a slight non-linearity in the relationship between judge leniency and released misconduct rates, with a slightly concave dashed curve for white defendants and a closer to linear dashed curve for Black defendants.

[18] If Z_i is independent of $(Y_i^*, D_{i1}, \ldots, D_{iJ}, R_i)$ given X_i and $E[Y_i^* \mid D_{ij} = 1, R_i = r, X_i] = \psi_{jr} + X_i'\gamma$, then $E[Y_i \mid R_i, Z_i, X_i, D_i = 1]$ is linear in $(W_i Z_{i1}, \ldots, W_i Z_{iJ}, Z_{i1}, \ldots, Z_{iJ}, X_i')'$, as in Equation (21). Analogously, if $E[D_{ij} \mid R_i = r, X_i] = \phi_{jr} + X_i'\beta$, under conditional random assignment $E[D_i \mid R_i, Z_i, X_i]$ is linear as in Equation (14). Appendix Table A6 relaxes the linearity assumption by estimating separate regression models for each NYC borough and averaging the resulting unwarranted disparities by a borough's share of cases. Reassuringly, we find similar (though less precise) estimates in in this specification.

[19] All standard errors in this and subsequent sections are obtained from a bootstrap procedure which accounts for the first-step estimation of the judge- and race-specific release rates and released misconduct rates.

Column 2 of Table 3 shows that the former nonlinearity translates to a somewhat lower estimate of white mean risk, at 0.319 (SE: 0.021), with a similar estimate of Black mean risk, at 0.394 (SE: 0.021). Near one, the non-parametric fit of Figure 2 coincides with the linear fit for white defendants and is above both the quadratic and linear fit for Black defendants, yielding mean risk estimates in column 3 of 0.346 (SE: 0.014) and 0.436 (SE: 0.016), respectively. The implied racial gap in risk—and thus the potential for OVB—rises with these more flexible extrapolations, to 7.5 percentage points in column 2 and 9.0 percentage points in column 3. We take the most flexible local linear extrapolation as our baseline specification for analyzing racial discrimination in NYC, which we show below gives the most conservative estimate of average discrimination. We also explore robustness to a wide range of alternative mean risk estimates below.

The extrapolations in Figure 2 yield accurate mean risk estimates when judge release rules are accurately parameterized or when there are many highly lenient judges. Appendix Figure A1 validates our extrapolations by plotting race-specific extrapolations of average predicted misconduct outcomes, among released defendants, in place of actual released misconduct averages in Figure 2. We first construct predicted misconduct outcomes $\hat{Y}_i^*$ using the fitted values from an OLS regression of actual pretrial misconduct Y_i^* on the controls in column 3 of Table 2 in the subsample of released defendants. We then plot estimates of $E[\hat{Y}_i^* \mid D_{ij} = 1, R_i = r]$ and $E[D_{ij} = 1 \mid R_i = r]$, constructed as in Figure 2, in Appendix Figure A1. Since $\hat{Y}_i^*$ can be computed for the entire sample, we also include the overall averages $E[\hat{Y}_i^* \mid, R_i = r]$ that are analogous to the race-specific ATEs of interest. Figure A1 shows that each of the linear, quadratic, and local linear extrapolations of predicted misconduct rates yields similar and accurate estimates of the overall actual averages. The 95 percent confidence intervals of the local linear extrapolations, for example, include the actual Black average and only narrowly exclude the actual white average. These results build confidence for the extrapolations of actual pretrial misconduct outcomes in this setting.[20]

Racial Discrimination

Panels B and C of Table 3 summarize the estimates of unwarranted racial disparities Δ_j given the corresponding ATE estimates in Panel A. These estimates are obtained from the sample analogue of Equation (9), noting that a judge's true negative rates can be written:

$$\delta_{jr}^T = E[D_{ij} \mid Y_i^* = 0, R_i = r] = (1 - E[Y_i^* \mid D_{ij} = 1, R_i = r]) \frac{E[D_{ij} \mid R_i = r]}{1 - \mu_r} \tag{22}$$

and similarly for her false negative rate δ_{jr}^F, while $\bar{\mu} = \mu_w p_w + \mu_b p_b$. We use the regression-adjusted estimates of $E[D_{ij} \mid R_i = r]$ and $E[Y_i^* \mid D_{ij} = 1, R_i = r]$ from Figure 2 and the sample share of Black defendants to complete this formula. Case-weighted averages of the resulting Δ_j estimates, reported in Panel B, estimate system-wide discrimination. We also compute empirical Bayes posteriors for individual Δ_j again via standard shrinkage procedures (Morris, 1983). Summary statistics for the judge-level prior distribution (estimated as in Figure 1) are reported in Panel C.

We find that approximately two-thirds of the system-wide release rate disparity between white and Black defendants in NYC is explained by racial discrimination, with about one-third explained

[20] Appendix Table A7 explores the sensitivity of our extrapolations to estimation error in judge release rates, which may attenuate their estimated relationship with released misconduct rates. We do so by first applying empirical Bayes shrinkage to the release rate estimates, separately by race. This exercise yields very similar results, suggesting negligible bias from first-step estimation error which is consistent with the fact that we observe many cases per judge ($\geq$100).

by unobserved differences in pretrial misconduct risk. The local linear extrapolations yield the most conservative estimate of system-wide discrimination in Table 3, implying that 62 percent (4.2 percentage points) of the case-weighted average disparity of 6.8 percentage points in Table 2 can be explained by discrimination. By comparison, both the linear and quadratic extrapolation-based estimates of race-specific mean risk imply that 79 percent (5.4 percentage points) of the average benchmarking disparity can be explained by racial discrimination. We thus find that unobservable differences in defendant risk can explain 21 to 38 percent (1.4 to 2.6 percentage points) of the average benchmarking disparity that remains after adjusting for court-by-time fixed effects.[21]

Appendix Table A8 illustrates how our rescaling approach yields this finding of significant racial discrimination in NYC bail decisions, following the simple numerical example in Appendix Table A5. We use the benchmark local linear estimates of mean risk to estimate the number of white and Black defendants with and without misconduct potential in column 1 of Panel A. In column 2, we combine these estimates with estimates of release and released misconduct rates adjusted by court-by-time fixed effects to compute the number of released defendants in each race and misconduct category, as in Equation (22). This calculation yields the case-weighted average observational disparity of 6.8 percentage points in column 6. In Panel B, we use the local linear estimates of mean risk to compute and apply the appropriate rescaling factor Ω_i. Our baseline estimates of average misconduct risk are $\mu_w = 0.346$ for white defendants and $\mu_b = 0.436$ for Black defendants. Combining these estimates with the share of white and Black defendants in our sample yields an overall average misconduct risk of $\bar{\mu} = 0.392$. Following Equations (17) and (18), these estimates yield a rescaling factor of $\Omega_i = \frac{1-0.392}{1-0.346} = 0.928$ for released white defendants with $Y_i = 0$, $\Omega_i = \frac{0.392}{0.436} = 0.901$ for released Black defendants with $Y_i = 1$, $\Omega_i = \frac{0.392}{0.346} = 1.137$ for released white defendants with $Y_i = 1$, and $\Omega_i = \frac{1-0.392}{1-0.436} = 1.077$ for released Black defendants with $Y_i = 0$. Thus the rescaling factor up-weights the release rates of risky white defendants and non-risky Black defendants (who are relatively less common) while down-weighting the release rates of non-risky white defendants and risky Black defendants (who are relatively more common). Applying these rescaling factors to the observational release rates yields a system-wide discrimination estimate of 4.2 percentage points, matching the estimate in Panel B of Table 3.

Figure 3 plots the full distribution of discrimination posteriors across individual bail judges, again using the most conservative local linear estimates of mean risk. For comparison, we also include the distribution of observed racial disparities from our benchmarking model that adjusts only for the court-by-time fixed effects. The former distribution is shifted evenly to the left of the latter distribution, consistent with nontrivial OVB across the judge-specific estimates. Around 62 percent of the judge-weighted average benchmarking disparity (4.2 percentage points, out of 6.6 percentage points) is found to be due to discrimination, the same as the case-weighted decomposition from Panel B of Table 3. The standard deviation of judge-specific unwarranted disparities remains large, at 3.7 percentage points, though it shrinks somewhat from the 4.0 percentage point standard deviation of observed release rate disparities. The clear majority of NYC judges have positive Δ_j, at 87.3 percent, though this share is also smaller than the 95.9 percent predicted by the benchmarking model. Panel C of Table 3 shows that these statistics are similar across different mean risk estimates.

[21] We can also use the decomposition (2) to compute the case-weighted disparity in true and false negative rates generating the overall 4.2 percentage point release rate disparity. From our baseline local linear extrapolation, we obtain an average $\delta^T_{jw} - \delta^T_{jb}$ of 2.7 (SE: 3.0) and an average $\delta^F_{jw} - \delta^F_{jb}$ of 6.6 (SE: 3.8). While noisy, these estimates suggest judges favor white defendants over Black defendants in both the $Y^*_i = 0$ and $Y^*_i = 1$ subpopulations.

Our estimates show that there are both statistically and economically significant inequalities in the release rate decisions of Black and white defendants with identical potential for pretrial misconduct. The most conservative estimate in Table 3, for example, implies that the unwarranted release rate gap could be closed if NYC judges released roughly 2,609 more Black defendants each year (or detained roughly 2,609 more white defendants). Using an estimate from Dobbie, Goldin and Yang (2018), releasing this many defendants would lead to around $78 million in recouped earnings and government benefits annually. We can also compare the average unwarranted disparity to other observed determinants of pretrial release. Table 2 shows, for example, that the most conservative 4.2 percentage point unwarranted disparity estimate corresponds to more than half of the decreased probability in release associated with having an additional pretrial arrest in the past year (-6.8 percentage points).

Robustness and Extensions

We verify the robustness of our main findings to several deviations from the baseline specification, exploring alternative estimates of mean risk, definitions of pretrial misconduct, classifications of pretrial release, and definitions of defendant race.

Mean Risk Estimates: The key inputs to our discrimination measure are race-specific estimates of mean misconduct risk, given in our baseline specification by a local linear extrapolation of the quasi-experimental variation in Figure 2. Appendix Figure A2 and Appendix Table A9 examine sensitivity to different ranges of these inputs, showing that our finding of pervasive racial discrimination does not depend on any particular extrapolation. Appendix Figure A2 first plots the range of system-wide unwarranted disparity that we would obtain from different pairs of white and Black mean risk inputs. The estimated level of discrimination against Black defendants generally decreases as the assumed value of Black misconduct risk increases, holding the value of white misconduct risk constant. Racial differences in misconduct risk would have to be extremely large, however, before we could conclude there is no discrimination against Black defendants. For example, at our baseline estimate of white mean risk (indicated by the dotted vertical line), the white-Black difference in misconduct risk would need to be more than 17 percentage points (88 percent) larger than our most conservative estimates in order for us to conclude that there is no discrimination against Black defendants.

Appendix Table A9 also shows that most of the mean risk inputs in Appendix Figure A2, and importantly all inputs which would imply no racial discrimination, can be ruled out by the observed quasi-experimental data. Panel A reports bounds on white and Black misconduct risk implied by the observed race-specific average released misconduct rate of judges with a given release rate, assuming that either none or all of the remaining detained defendants of each race have misconduct potential (see Appendix B.5 for details). Panels B and C report corresponding bounds on the unwarranted disparity statistics by finding the pair of mean risk estimates which minimize and maximize each statistic in these ranges. The bounds on each statistic widen as a lower release rate is used, since a wider range of mean risk estimates are consistent with the increasingly selected released misconduct rates. Nevertheless, we continue to find economically significant levels of racial discrimination in each column, even when a relatively low release rate of 0.80 is used to construct the bounds.

Finally, Appendix Table A10 shows that we obtain similar mean risk estimates when extrapolating released misconduct rates that adjust for defendant and case observables, as in column 3 of Table 2. Panel A shows, for example, that the local linear extrapolation yields white and Black mean risk

estimates of 0.352 and 0.423, respectively, compared to the 0.346 and 0.436 in our baseline Table 3 which does not adjust for defendant and case observables. We obtain slightly smaller unwarranted disparity estimates in Panel B of Appendix Table A10, suggesting that some of the unwarranted disparity we find in Table 3 is driven by discrimination on seemingly race-neutral characteristics. Benchmarking comparisons that adjust for these characteristics thus remove some of the drivers of discrimination in NYC bail decisions, as well as some amount of omitted variables bias. However, we do not observe all of the non-race characteristics that judges may base decisions on. Therefore, this analysis cannot tell us the fraction of discrimination which arises through discrimination on non-race characteristics, as there could be further discrimination on unobserved characteristics that we cannot detect in Appendix Table A10.

Misconduct Outcome: Our baseline measure of racial discrimination assumes that the sole legal objective of bail judges is to target pretrial misconduct, and not other objectives or outcomes. When the legal objective of judges is misspecified, our estimates may suffer from what Kleinberg et al. (2017) refer to as "omitted payoff bias." Such bias may arise when, for example, bail judges consider new crime to be more important than a failure to appear, or if they only target new violent crime. We explore the empirical relevance of omitted payoff bias in Appendix Table A11, which presents estimates given these different definitions of the judge's legal objective. We find similar results when using a measure of pretrial misconduct that only includes FTA (column 2 of Appendix Table A11) or only includes new arrests (column 3 of Appendix Table A11). We also find a slightly higher case-weighted average unwarranted disparity, at 6.8 percentage points, when using a measure of pretrial misconduct that only includes new arrests for a violent crime (column 4 of Appendix Table A11). These results are consistent with Kleinberg et al. (2017) and Arnold, Dobbie and Yang (2018), who find similar evidence of prediction errors and racial bias in bail decisions, respectively, using different measures of the pretrial misconduct outcome.

A related concern is that measurement error in the judge's legal objective is systematically correlated with race. This could be an issue if, for example, judges seek to minimize all new crime, not just new crime that results in an arrest, and if the police are more likely to rearrest Black defendants conditional on having committed a new crime. Gelman, Fagan and Kiss (2007), for example, find that the NYC Stop, Question, and Frisk program disproportionately targeted minority residents. With discriminatory policing, we will tend to overestimate the misconduct risk for Black defendants compared to white defendants and underestimate the total amount of racial discrimination in bail decisions. It is therefore possible that our estimates reflect a lower bound on the true amount of racial discrimination in NYC, at least under the plausible assumption that the police are more likely to rearrest Black defendants conditional on having committed a new crime. Reassuringly, column 2 of Appendix Table A11 shows a similar level of discrimination when we measure pretrial misconduct using just FTA, which is less subject to this measurement concern.

Release Decision: Our baseline specification abstracts away from the fact that bail judges may set different levels of monetary bail, taking into account a defendant's ability to pay, by specifying the judge's decision as a binary release indicator. One possibility is that the discrimination we find is partly driven by judges over-predicting the relative ability of Black defendants to pay cash bail, causing fewer Black defendants to be released than white defendants of identical misconduct risk. We explore racial differences in the ability to pay cash bail in Appendix Table A12, which replaces our

baseline definition of the judge's release decision with an indicator for the judge releasing a defendant on recognizance, without setting cash bail. We find very similar results with this new specification, with racial discrimination explaining about 55 percent (3.2 percentage points) of the court-by-time adjusted white-Black ROR rate difference of 5.8 percentage points. These results suggest that the racial discrimination we find in bail decisions is not driven by judges over-predicting the relative ability of Black defendants to pay cash bail, which is consistent with the fact that the vast majority of released white and Black defendants are released on recognizance (see Table 1).

Defendant Race: A final consideration is our categorization of defendant race. We categorize defendants as either white (including both non-Hispanic and Hispanic white individuals) or Black (including both non-Hispanic and Hispanic Black individuals), but judges may also discriminate against Hispanic white defendants. We explore this possibility in Appendix Table A13, which presents estimates with defendants categorized as either non-Hispanic white or any racial minority (including Hispanic white individuals and both non-Hispanic and Hispanic Black individuals). Under this alternative categorization, we find larger estimates of case-weighted average unwarranted disparity, for example, 11.2 percentage points for the local linear extrapolation in column 3. One important caveat to these results is that the extrapolation for white defendants in this specification relies on much fewer observations, which leads to less consistent results across specifications. Therefore, these results should be interpreted with caution, as our procedure relies on consistent estimates of mean risk that may be more difficult to obtain with fewer observations.

Defendant Heterogeneity

Appendix Table A14 explores heterogeneity in racial discrimination across defendants with different observable characteristics. We report estimates using a conditional version of our baseline local linear approach that restricts to defendants with a particular criminal record or charge.[22] Panel A shows that the average misconduct risk within different subgroups of Black defendants is consistently higher than the average misconduct risk of white defendants in the same subgroup, with point estimates for the racial difference in average misconduct risk ranging from 3.5 percentage points for defendants charged with a DUI and 8.3 percentage points for defendants charged with a misdemeanor, to 11.5 percentage points for defendants charged with a drug offense, 12.1 percentage points for defendants charged with a property offense, 14.3 percentage points for defendants charged with a felony, and 17.4 percentage points for defendants charged with a violent offense. The common finding of a positive racial gap in average misconduct risk implies that observed release rate disparities will suffer from omitted variables bias in each subgroup.

Panel B of Appendix Table A14 shows that we find discrimination against Black defendants in each subgroup, with point estimates for the extent of discrimination ranging from 1.0 percentage points for defendants charged with a property offense and 2.4 percentage points for defendants charged with a DUI and defendants without a prior criminal charge, to 3.0 percentage points for defendants charged with a felony, 4.6 percentage points for defendants charged with a misdemeanor, 5.5 percentage points for defendants charged with a drug offense, and 10.7 percentage points for defendants charged with a violent offense. The estimates are generally precisely estimated, with the exception of felony offenses

[22] We require that judges observe at least 25 cases involving defendants with the indicated criminal record or charge in each specification, meaning that these results include fewer judges than our main results.

and violent offenses where we have large standard errors. Panel C shows a similar pattern of results in the average unwarranted disparity across judges, with significant discrimination against Black defendants in each subgroup and larger but less precise estimates for felonies and violent offenses.

Judge Heterogeneity

Finally, Table 4 explores the heterogeneity in racial discrimination across different judges in our sample. Columns 1-5 report OLS estimates of the unwarranted disparity posteriors on indicators for whether a judge is newly appointed during our sample period, exhibits above-average leniency, or has an above-median share of Black defendants (as measured before the adjustment for court-by-time fixed effects, which makes Black defendant shares balanced across judges). We also include indicators for what county courtroom the judge hears most cases in. Columns 6-7 investigate the persistence of our discrimination measure over time by computing separate unwarranted disparity posteriors in the first and second half of cases that each judge sees in our sample period, recomputing the race-specific mean risk estimates in each half, and estimating OLS regressions of current unwarranted disparity posteriors on lagged unwarranted disparity posteriors and judge observables. In both sets of analyses, regressions of discrimination posteriors on judge observables can be interpreted through the posterior average effect framework of Bonhomme and Weidner (2020). We weight these regressions by estimates of the inverse posterior variance of the unwarranted disparities, with similar results obtained from weighting by judge caseload.

Columns 1-5 of Table 4 shows significantly lower levels of discrimination among newly appointed judges, more lenient judges, and judges with a higher share of Black defendants. Judges who are newly appointed in our sample have 1.6 percentage point lower unwarranted disparities on average, while judges with above-average leniency have 0.9 percentage point lower unwarranted disparities. Judges assigned an above-median share of Black defendants have 1.2 percentage point lower unwarranted disparities. We also find that judges who primarily see cases in the Manhattan, Queens, and Richmond county courtrooms tend to exhibit higher levels of discrimination, while those who primarily see cases in Brooklyn (the omitted reference category) and the Bronx have lower levels of discrimination. We find, for example, that unwarranted disparities are 3.6 percentage points higher for Manhattan judges compared to Brooklyn judges. Together, the observable judge characteristics available in our data explain about 41 percent of the variation in the unwarranted disparity posteriors, with the courtroom indicators alone explaining about 35 percent of the variation in unwarranted disparities.

Columns 6-7 of Table 4 show that the judge-specific discrimination estimates are highly correlated over time, with an autoregressive coefficient of 0.56. Lagged unwarranted disparities alone explain about 25 percent of the variation in current unwarranted disparities, with the lagged disparity and observable judge characteristics explaining about 35 percent. We also note that the average unwarranted disparity in the second half of judge cases is somewhat larger, at 5.6 percentage points, suggesting that discrimination may increase with judge experience.

Taken together, the results from this section robustly show that there is substantial racial discrimination in NYC bail decisions, both on average and for most defendants and judges, and that judge-specific estimates of discrimination are both predicted by observable characteristics and correlated over time. However, these results do not speak to whether such discrimination is driven by racial bias or statistical discrimination, nor whether we can reliably target and potentially reduce racial discrimination using existing data. We next consider a framework to answer these questions.

6 MTE Estimates of Bias and Statistical Discrimination

6.1 Methods

We develop and estimate a hierarchical marginal treatment effects (MTE) model that imposes additional structure on our quasi-experimental variation to investigate whether discrimination in bail decisions is driven by racial bias or statistical discrimination, and to conduct policy simulations. Building on the illustrative model in Section 3.2, we suppose that judges base release decisions on noisy signals of true misconduct potential. We allow for judge- and race-specific risk preferences and signal quality, with the latter allowing heterogeneous race-specific predictive skill across judges (in violation of conventional first-stage monotonicity). The model implies a distribution of judge- and race-specific MTE curves that can be used to estimate racial bias at the margin of release, as well as to measure racial differences in average risk or signal quality that can generate statistical discrimination.

As before, we model judge risk signals as $\nu_{ij} = Y_i^* + \eta_{ij}$, where $\eta_{ij} \mid Y_i^*, (R_i = r) \sim N(0, \sigma_{jr}^2)$ denotes the noise in judge j's risk signals for defendants of race r. Signal quality is given by the inverse standard deviation of noise, $\tau_{jr} = 1/\sigma_{jr}$, such that higher τ_{jr} corresponds to more precise risk signals. Judges with higher τ_{jr} can be thought of as having a richer information set or as being more skilled at inferring true misconduct potential from a common information set. Judges combine these race-specific signals τ_{jr} with potentially biased prior beliefs $\tilde{\mu}_{jr}$ of mean misconduct risk μ_r for each race r and an understanding of the signal-generating process. The judges' risk posteriors $p_j(\nu_{ij}; R_i)$ are therefore potentially biased solutions to the binary classification problem of whether defendant i would fail to appear or be rearrested for a new crime if released ($Y_i^* = 1$), given the individual's race r and noisy misconduct signal ν_{ij}. Appendix B.2 derives these posterior functions and shows that they are strictly increasing in the risk signal. Given release benefits π_{jr}, the release decisions of each risk-neutral judge therefore follow a signal-threshold rule of:

$$D_{ij} = \mathbf{1}[\pi_{jR_i} \geq p_j(\nu_{ij}; R_i)] = \mathbf{1}[\kappa_{jR_i} \geq Y_i^* + \eta_{ij}] \tag{23}$$

where $\kappa_{jr} = p_j^{-1}(\pi_{jr}; r)$ is an implicit function of judge j's release benefit π_{jr}, subjective risk belief $\tilde{\mu}_{jr}$, and risk signal quality τ_{jr} for defendants of race r. Appendix B.6 shows that when judges respond to misconduct risk, such that $\delta_{jr}^T > \delta_{jr}^F$, there exists a signal threshold κ_{jr} and signal quality $\tau_{jr} > 0$ which rationalize the reduced-form true and false negative rates. Absent further restrictions, this model is thus without observational loss so long as judge release decisions are better-than-random.

When known for each race, a judge's risk threshold κ_{jr} and signal quality τ_{jr} can be used to characterize the extent of racial bias in release decisions. As discussed in Section 3.2, with accurate beliefs the average misconduct outcomes at the margin of pretrial release capture the race-specific release benefits $\pi_{jr} = E[Y_i^* \mid p_j(\nu_{ij}; r) = \pi_{jr}] = E[Y_i^* \mid Y_i^* + \eta_{ij} = \kappa_{jr}]$, which can be used to compute racial bias for judge j.[23] These marginal released outcomes are known functions of κ_{jr} and τ_{jr}, and represent marginal treatment effects (of release on pretrial misconduct) for defendants at the margin of release. Arnold, Dobbie and Yang (2018) use marginal released outcomes to test for racial bias among quasi-randomly assigned bail judges under an assumption of first-stage monotonicity, which

[23] Here, as before, we consider racial bias due either to racial preferences or biased beliefs. Appendix B.2 shows how differences in release benefits and prior risk beliefs are observationally equivalent in this model. Both terms enter the κ_{jr} multiplicatively, such that for any $\kappa \in \mathbb{R}$ and $\tau_{jr} > 0$ there exists a set of π_{jr} and $\tilde{\mu}_{jr}$ (each ranging from 0 to 1) with $\kappa_{jr} = \kappa$. This equivalence reflects the general difficulty of disentangling racial bias due to biased beliefs (as in Bordalo et al., 2016) from racial bias due to taste-based bias (as in Becker, 1957).

here imposes $\eta_{ij} = \eta_i$ (and thus $\tau_{jr} = \tau_r$) to be common to all judges, such that judges act as though there is a common ordering of defendants (of each race) with regards to their appropriateness for release. Under this monotonicity restriction, the race-specific marginal released outcomes needed to test for bias can be estimated with conventional MTE estimation methods.

Our first insight here is that knowledge of κ_{jr} and τ_{jr} can also be used to measure the extent of statistical discrimination. As discussed in Section 3.2, statistical discrimination arises when judges act on risk predictions that are affected by racial differences in either mean misconduct risk μ_r or signal quality τ_{jr}. Mean risk for each race r is given by integrating the marginal released outcome (or MTE) curve $\mu_{jr}(\kappa) = E[Y_i^* \mid Y_i^* + \eta_{ij} = \kappa]$ for each judge j over the distribution of her risk signals. The slopes of these curves capture the quality of a judge's risk signals. Relatively more precise signals for white defendants relative to Black defendants will, for example, lead to a steeper-sloping $\mu_{jw}(\kappa)$ relative to $\mu_{jb}(\kappa)$. More generally, the judge- and race-specific MTE curves $\mu_{jr}(\kappa)$ can be used to calculate the extent of racial discrimination in counterfactuals calculations where a judge's release rates are set to equalize marginal released outcomes and eliminate racial bias.

Our second insight is that we can estimate the key κ_{jr} and τ_{jr} parameters of the model without imposing a strong assumption of first-stage monotonicity. By restricting $\eta_{ij} = \eta_i$ and thus $\tau_{jr} = \tau_r$, monotonicity can be understood to restrict the MTE curves $\mu_{jr}(\cdot)$ to be common across judges for each race r, such that variation in judge release rates only reflects differences in risk thresholds κ_{jr}. An implication of this restriction is that, absent estimation error, the race-specific release rates $E[D_{ij} \mid R_i = r]$ and released misconduct rates $E[Y_i^* \mid D_{ij} = 1, R_i = r]$ plotted in Figure 2 will lie on a single curve determined by the common signal quality τ_r and mean risk μ_r. Given the large number of cases per judge in NYC, the sizable dispersion we see in the figure is thus indicative of monotonicity violations. Frandsen, Lefgren and Leslie (2019) formalize this logic by building on similar tests of monotonicity in the context of quasi-randomly assigned judges (Mueller-Smith, 2015; Norris, 2019) and elsewhere (Kitagawa, 2015). Appendix Table A15 shows that applying the Frandsen, Lefgren and Leslie (2019) test to our data yields decisive rejections, in both samples of white and Black defendants, suggesting conventional monotonicity is unlikely to hold across NYC bail judges.

We therefore substitute the conventional monotonicity restriction with an alternative parameterization of heterogeneity in judge skill, permitting a distribution of MTE curves $\mu_{jr}(\cdot)$ across judges rather than restricting $\mu_{jr}(\cdot) = \mu_r(\cdot)$ across all judges j. We specify the signal quality parameters τ_{jr} as being log-normally distributed (imposing the domain restriction of $\tau_{jr} > 0$), jointly with the signal thresholds κ_{jr}: $\ln \tau_{jr} \sim N(\alpha_r, \psi_r^2)$ and $\kappa_{jr} \sim N(\gamma_r, \delta_r^2)$ with non-zero correlations allowed across j and r. Appendix B.6 shows how this hierarchical approach can be viewed as parameterizing differences in how judges weigh different defendant characteristics, such as demeanor or prior arrest record.

We estimate the hyperparameters governing the distributions of judge-specific MTE curves by a simulated minimum distance (SMD) procedure that matches moments of the quasi-experimental release rate and released misconduct rate variation in Figure 2. This procedure, detailed in Appendix B.7, first estimates race-specific curves-of-best-fit through race-specific release and released misconduct rates (as in Section 5.2). We then match the estimated intercept, slope, and curvature of these curves-of-best-fit, as well as the residual variation in first-step estimates, to the corresponding moments of simulated quasi-experimental data drawn from different parameterizations of the hierarchical MTE model. Finally, we use the SMD estimates to compute empirical Bayes posteriors of the marginal released outcomes and signal quality of each judge and race given the hyperparameter estimates and

observed quasi-experimental data.

Figure 4 builds intuition for the SMD estimation procedure by showing how differences in key hyperparameters manifest in the quasi-experimental data. We construct this figure by first simulating draws of κ_{jr} and τ_{jr} for a given race r across a large population of judges j with arbitrarily varying leniency. We then plot the implied distribution of judge release rates $E[D_{ij} \mid R_i = r]$ and released misconduct rates $E[Y_i^* \mid D_{ij} = 1, R_i = r]$, abstracting away from first-step estimation error. Panels A and B set the variance of signal quality across judges to zero, satisfying the usual first-stage monotonicity restriction and ensuring that the judge moments fall on a common frontier.[24] Panels C and D then relax monotonicity by allowing signal quality to vary across judges.

Panel A of Figure 4 shows how differences in mean misconduct risk μ_r lead to differences in the vertical intercept of these curves at one, or (per the discussion in Section 5.1) the release rate of a hypothetical supremely lenient judge. These vertical intercepts correspond to model-based extrapolations of the quasi-experimental data, in contrast to the data-driven extrapolation used previously in Section 5. Panel B further shows how differences in mean signal quality lead to different slopes of the model-implied curves, with higher τ_r resulting in a steeper relationship between the share of defendants that a judge releases and the extent of pretrial misconduct among the released. When we relax first-stage monotonicity in Panels C and D, the quasi-experimental variation no longer falls on a common frontier (even without estimation error). Panel C shows that a higher variance in signal quality manifests as more dispersion in released misconduct rates among judges with similar release rates. Such dispersion generates rejections of the monotonicity tests developed by Frandsen, Lefgren and Leslie (2019) and others. Finally, Panel D shows that the trend in this distribution of points becomes more nonlinear when judge signal quality is more highly correlated with judge leniency.

6.2 Results

Racial Bias and Statistical Discrimination

Table 5 reports SMD estimates of mean misconduct risk μ_r, the average misconduct outcomes for marginally released defendants $\mu_{jr}(\kappa_{jr})$ across judges, and the average judge signal quality τ_{jr}, with the underlying hierarchical MTE model hyperparameter estimates reported in Appendix Table A16. The average difference in marginal misconduct outcomes between white and Black defendants captures the overall extent of racial bias, while differences in either mean risk or signal quality by race capture statistical discrimination. Columns 1-3 of Table 5 report estimates under the conventional first-stage monotonicity restriction that signal quality for defendants of a given race is constant across judges. Columns 4-6 relax this restriction, allowing judges to have different rankings of defendant appropriateness for pretrial release.[25]

[24] By restricting $\tau_{jr} = \tau_r$, we still allow for random violations of monotonicity in the sense of $\eta_{ij} \neq \eta_{ik}$ for $j \neq k$, so long as η_{ij} and η_{ik} have the same variance. This assumption is akin to the notion of "average monotonicity" in Frandsen, Lefgren and Leslie (2019). Similarly, when we allow τ_{jr} to differ, two judges with the same signal quality may nevertheless have different rankings η_{ij} over defendants.

[25] The estimates in columns 1-3 of Table 5 are derived from the hyperparameter estimates in columns 1 and 4 of Appendix Table A16, while columns 4-6 of Table 5 come from columns 2 and 5 of Appendix Table A16. The latter assumes log signal quality and release thresholds are uncorrelated. A richer model that allows for such correlation is estimated in columns 3 and 6 of Appendix Table A16. This model produces estimates that are very similar to columns 2 and 5 but also considerably less precise. We therefore take the uncorrelated model as our baseline in Table 5. We note that our baseline model still allows for correlation between judge signal quality and marginal released outcomes, which we find to be large in Table 5. Appendix Figure A4 shows how these model hyperparameters fit the quasi-experimental variation by plotting the model-implied average released misconduct rate across races and judges of different leniencies,

In both sets of model estimates, we find evidence of both racial bias and statistical discrimination, with the latter coming from a higher level of average risk (that exacerbates discrimination) and less precise risk signals (that alleviates discrimination) for Black defendants. Columns 4-6 of Table 5 show, for example, that the expected misconduct rate of typical white defendants at the margin of pretrial release is 0.651 (SE: 0.033), compared to 0.576 (SE: 0.021) for Black defendants. The difference in these mean marginally released outcomes is a statistically significant 7.4 percentage points (SE: 3.8), indicating the existence of racial bias at the margin of release. Table 5 further shows considerable scope for statistical discrimination. First, the model estimates confirm the finding in Section 5.2 that mean risk is lower among white defendants than Black defendants. In columns 4 and 5, this difference in mean misconduct risk is 5.0 percentage points, about half the size of the 9.0 percentage point difference from our most conservative local linear extrapolation in Table 3 but similar to the 6.0 percentage point gap from our simple linear extrapolation. Second, we find that the typical judge acts on higher-quality risk signals for white defendants than for Black defendants. Columns 4 and 5 of Table 5 report an average signal quality of 1.385 (SE: 0.104) for white defendants and 0.970 (SE: 0.073) for Black defendants, implying that the typical noise in Black risk signals is roughly 30 percent more dispersed. Per the discussion of Figure 4, this result is consistent with the white line-of-best-fit from Figure 2 being somewhat steeper than the Black line-of-best-fit.[26] With a majority of white and Black defendants released, higher white signal quality is likely to offset racial discrimination against Black defendants arising from other channels (see Section 3.2). Together, the racial differences in mean risk and signal quality imply that analyses of racial bias alone (as in Arnold, Dobbie and Yang (2018) and Marx (2018)) would omit an important source of discrimination in this setting.

Table 5 further suggests that the conventional first-stage monotonicity restriction is inconsistent with judge behavior in this setting. We find significant variation in judge signal quality when we relax this restriction and allow judges to have different rankings of defendant appropriateness for pretrial release in columns 4-6, with standard deviations of 0.196 (SE: 0.038) for white defendant signal quality and 0.163 (SE: 0.017) for Black defendant signal quality. This variation in judge skill is highly correlated with variation in judge release preferences (which we also find to be sizable), with covariances between judge signal quality and marginal released outcomes of 0.013 for white defendants and 0.007 for Black defendants (implying respective correlation coefficients of 0.83 and 0.67). While point estimates of the mean parameters with and without conventional monotonicity are qualitatively similar, the precision is higher without. The standard error on average racial bias, for example, falls by 17 percent from column 3 to column 6. These precision gains also suggest that the model without monotonicity provides a better fit to the quasi-experimental data, consistent with a visual analysis of Figure 2 and the formal tests in Appendix Table A15. At the same time, the similarity of the estimates across the columns of Table 5 suggests that imposing an invalid assumption of monotonicity in this setting does not qualitatively affect the results.

Appendix Table A19 uses the unrestricted model to quantify the joint role of racial bias and statistical discrimination in driving racial discrimination in NYC bail decisions. Column 1 summarizes

along with the estimates of release rates and released misconduct rates from Figure 2. Both model-implied curves-of-best-fit are approximately linear, with slight upward curvature and a more steeply sloping curve for white defendants.

[26] The mean signal quality estimates in Table 5 suggest that the typical NYC judge predicts misconduct risk with considerable accuracy for both races. In terms of the model, a τ_{jr} of 1.385 (0.970) yields a receiver operating characteristic curve with an area under the curve (AUC) statistic of 0.835 (0.753) for white (Black) defendants. By comparison, Kleinberg et al. (2017) obtain an AUC of 0.707 with a machine learning algorithm trained on FTA outcomes among released NYC defendants. Simpler logit models which use the observables in column 3 of Table 2 to predict Y_i^* among released defendants in our sample have AUCs of around 0.66 (0.65) for white (Black) individuals.

the baseline degree of discrimination, racial bias, and differences in signal quality. The model-based estimate of average unwarranted disparity, at 4.7 percentage points, is similar but somewhat higher than our most conservative estimate in Table 3.[27] Column 2 shows that average racial discrimination significantly declines when judge leniency is counterfactually raised or lowered to equalize marginal released outcomes across white and Black defendants (with Panel A generally raising Black release rates and Panel B generally lowering white release rates). The average unwarranted disparity falls from 4.7 percentage points to -4.2 percentage points in Panel A and -0.6 percentage points in Panel B. This result shows that, absent racial bias, the average unwarranted disparity is reversed, with white defendants becoming less likely to be released than Black defendants of identical misconduct potential. As expected, columns 3 and 4 show that this reversal is driven by the relatively higher signal quality for white defendants. Equalizing signal quality across races for each judge, with and without racial bias, again results in average racial discrimination against Black defendants. The remaining statistical discrimination solely due to mean risk differences in column 4 yields a mean unwarranted disparity of 3.9 percentage points when Black leniency and signal quality are counterfactually set, and a mean unwarranted disparity of 6.2 percentage points when adjusting the corresponding white parameters.[28]

Judge Heterogeneity

Appendix Tables A21–A22 explore variation in empirical Bayes posteriors of racial bias and signal quality differences, following our analysis of the unwarranted disparity posteriors in Section 5.2. We again report OLS estimates of the indicated posteriors on indicators for whether a judge is newly appointed during our sample period, has above-average leniency, has an above-median share of Black defendants, and for what county courtroom the judge hears most cases in. We again weight these regressions by estimates of the inverse posterior variance of the outcome variables, with very similar results again obtained when weighting by judge caseload.[29]

In Appendix Table A21, we find significantly lower levels of racial bias among newly appointed judges and less lenient judges. Courtroom indicators are also highly predictive. Together, the observable judge characteristics explain about 40 percent of the variation in racial bias across judges, with the courtroom indicators alone explaining 33 percent. We also find a strong relationship between racial bias and overall discrimination, with our discrimination measure explaining 65 percent of the variation in the judge-specific bias.

In Appendix Table A22, we find a relatively smaller racial gap in signal quality among newly appointed judges. Here, judge leniency and whether the judge has an above-median Black defendant share are not significant predictors of judge-specific signal quality by race. Courtroom indicators and other observable characteristics of the judges again explain much of the variation in signal quality

[27] All conclusions in Section 5.2, including the fraction of discriminatory NYC judges and heterogeneity results, continue to hold with the MTE model estimates of μ_r (see Appendix Figure A3 and Appendix Table A17).

[28] Appendix Table A20 uses the structure of the model to investigate alternative measures of racial disparity in NYC bail decisions. The first row reports the implied average release rates of white and Black defendants holding fixed misconduct potential Y_i^*, the difference in which gives our model-based estimate of system-wide unwarranted disparity (4.7 percentage points). The second row instead holds fixed the implied distribution of misconduct signals ν_{ij}, with the conditional white-Black release rate gap yielding a measure of "race-blindness." With an insignificant gap of -0.2 percentage points, the model estimates suggest virtually all discrimination in NYC bail decisions is through seemingly race-neutral characteristics. Finally, the third row instead holds fixed the implied distribution of misconduct posteriors $p_j(\nu_{ij}, R_i)$ instead of true misconduct potential Y_i^*. This result yields a measure of average racial bias, with the model implying a release rate gap of 7.1 percentage points due to the release thresholds for white and Black defendants.

[29] Estimates exploring variation of racial bias and signal quality differences by defendant heterogeneity, following Appendix Table A14, are imprecise but suggest qualitatively similar findings of non-zero racial bias and a positive white-Black gap in signal quality.

differences, with 38 percent of the variation explained when we include all judge observables. We find a stronger relationship between signal quality differences and overall discrimination than between racial bias and discrimination, with our discrimination measure explaining 74 percent of the variation in the judge-specific signal quality.

7 Policy Simulations

Lastly, we use our hierarchical MTE model estimates to investigate whether racial discrimination can be reliably targeted and potentially reduced with existing data. The model-free analysis in Section 5 shows that judge-specific unwarranted disparities are relatively stable over time, suggesting that identifying and targeting highly discriminatory judges for an appropriate intervention could help reduce future discrimination. This analysis also shows that approximately one-third of the observed release rate disparity between white and Black defendants is explained by unobserved differences in misconduct risk, suggesting that observational regressions may also be useful for targeting judge-specific discrimination even in the absence of our quasi-experimental analysis. By linking unobserved differences in misconduct risk, racial bias, and statistical discrimination in the release decisions of each judge, the hierarchical MTE model provides the necessary structure to simulate the effects of reducing racial discrimination using existing observational and quasi-experimental data. We focus on the more general question of whether discrimination can be reliably targeted using existing data, abstracting away from the legal status of any particular policy reform.

Table 6 summarizes simulations that target both unwarranted disparity posteriors (columns 2 and 3) and observational disparities (columns 4 and 5). The simulations suppose that individual bail judges can be subjected to race-specific release rate quotas that eliminate racial disparities, as estimated by a policymaker using either an observational or quasi-experimental analysis. The simulation based on the unwarranted disparity posteriors gauges the reliability of the individual predictions given the noise in our estimation procedure. The simulation based on observational disparities further tests whether conventional benchmarking regressions may be useful for targeting discrimination despite OVB. To simulate both sets of policies, we redraw all judge-specific parameters for each race from the estimated hierarchical MTE model 250 times, along with draws of appropriate estimation error. We use these to simulate 250 draws of the quasi-experimental variation plotted in Figure 2. We then re-estimate the MTE model in each draw and compute empirical Bayes posteriors, as in our analysis of the true data. Finally, we force all or a subset of simulated judges to adjust their race-specific leniencies to the point where their racial disparities are expected to be eliminated given the simulated model estimates and posteriors. Panel A simulates closing the targeted disparities for all judges, while Panel B simulates closing the targeted disparities only for judges in the top quintile of the estimated disparities.[30]

The simulations suggest that racial discrimination can be reliably targeted using our estimated unwarranted disparity posteriors, despite estimation error. Targeting the disparities of all judges using the unwarranted disparity posteriors results in the virtual elimination of racial discrimination (columns 2 and 3 of Table 6, Panel A), while only targeting judges in the top quintile results in a 36 percent reduction in the average level of discrimination (columns 2 and 3 of Panel B). These

[30] Column 1 of Table 6 shows baseline simulated averages of unwarranted disparity, observational disparity, and racial bias. Column 1 reports an average unwarranted disparity of 4.7 percentage points, which is a bit larger than the 4.2 percentage point average unwarranted disparity found with our local linear extrapolation in Section 5 due to the difference in model mean risk estimates.

simulated reductions are essentially unchanged when the targeted judges are forced to increase their leniency (typically for Black defendants) in column 2 or decrease their leniency (typically for white defendants) in column 3. The average standard deviation of unwarranted disparity across judges, reported in brackets, is also reduced from around 3.7 percentage points to 2.0 percentage points in column 2 and 2.6 percentage points in column 3. Observational release rate disparities still remain when eliminating discrimination, however, as the higher level of mean risk for Black defendants leads to OVB in the policy target.

Targeting judges with observational comparisons can also reduce discrimination, as the observed release rate disparities are highly correlated with the unwarranted disparity posteriors. Appendix Figure A5 shows, for example, that we obtain a high "forecast" coefficient of 0.903 (SE: 0.010) from regressing estimated judge-specific unwarranted disparity posteriors on observational disparity posteriors, along with a very high R-squared of 0.968. Consequently, we find in Table 6 that targeting all judges with simulated observational disparity posteriors reduces average unwarranted disparity by 6.4 percentage points in column 4 and 6.6 percentage points in column 5. The resulting average unwarranted disparity estimates of -1.7 and -1.9 percentage points reflects the fact that the level of observed disparities is too high on average because of OVB. When targeting just the observational disparity posteriors in the top quintile of judges, the average unwarranted disparity is reduced by 45 percent but not reversed (columns 4 and 5 of Panel B). This finding, that observational benchmarking regressions can be useful for monitoring and targeting racial discrimination despite OVB, mirrors a result in the education and healthcare setting on the utility of biased observational measures of school and hospital quality (e.g., Angrist et al., 2017; Hull, 2020). There, as here, observational rankings prove to be highly predictive of policy-relevant parameters despite non-zero bias.[31]

8 Conclusion

Large racial disparities exist at every stage of the criminal justice system, but it is unclear whether these disparities reflect racial bias, statistical discrimination, or omitted variables bias. This paper shows that racial discrimination in bail decisions can be measured, regardless of its source, using observational comparisons of white and Black release rates that are rescaled with quasi-experimental estimates of average white and Black misconduct risk. Our most conservative estimates from NYC show that approximately two-thirds of the observed racial disparity in release decisions is due to racial discrimination, with around one-third due to unobserved racial differences in misconduct risk. Leveraging a novel hierarchical MTE model, we show that this discrimination is driven by both racial bias and statistical discrimination, with the latter due to a higher level of average risk (that exacerbates discrimination) and less precise risk signals (that offsets discrimination) for Black defendants. Outcome-based tests of racial bias therefore omit an important source of racial discrimination in NYC bail decisions, and cannot be used to rule out all possible violations of U.S. anti-discrimination law.

We conclude by noting that the methods we develop to study racial discrimination in bail decisions may prove useful for measuring unwarranted disparities in several other high-stakes settings,

[31] Our simulations also highlight the impossibility of simultaneously eliminating racial discrimination (on average) and racial bias (at the margin) when either mean misconduct risk or the risk signal quality differ for white and Black defendants (Kleinberg, Mullainathan and Raghavan, 2017). The simulation based on the unwarranted disparity posteriors, for example, results in non-zero racial bias against Black defendants of between 1.3 and 3.9 percentage points at the margin of release.

both within and outside of the criminal justice system. One key requirement is the quasi-random assignment of decision-makers, such as judges, police officers, employers, government benefits examiners, or medical providers. A second requirement is that the objective of these decision-makers is both known and well-measured among the subset of individuals that the decision-maker endogenously selects. Mapping these settings to the quasi-experimental methods in this paper can help distinguish between different explanations for observed racial disparities and form appropriate policy responses.

References

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

Aigner, Dennis, and Glen Cain. 1977. "Statistical Theories of Discrimination in Labor Markets." *Industrial and Labor Relations Review*, 30(2): 157–187.

Andrews, Donald, and Marcia Schafgans. 1998. "Semiparametric Estimation of the Intercept of a Sample Selection Model." *Review of Economic Studies*, 65(3): 497–517.

Angrist, Joshua, Peter Hull, Parag Pathak, and Christopher Walters. 2017. "Leveraging Lotteries for School Value-Added: Testing and Estimation." *Quarterly Journal of Economics*, 132(2): 871–919.

Antonovics, Kate, and Brian Knight. 2009. "A New Look at Racial Profiling: Evidence from the Boston Police Department." *Review of Economics and Statistics*, 91(1): 163–177.

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

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

Arrow, Kenneth J. 1973. "The Theory of Discrimination." In *Discrimination in Labor Markets.* , ed. Orley Ashenfelter and Albert Rees, 3–33. Princeton, NJ: Princeton University Press.

Ayres, Ian. 2010. "Testing for Discrimination and the Problem of Included Variable Bias." *Yale Law School Mimeo.*

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

Berk, Richard, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2018. "Fairness in Criminal Justice Risk Assessments: The State of the Art." *Sociological Methods & Research*, 1–42.

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

Bonhomme, Stephane, and Martin Weidner. 2020. "Posterior Average Effects." *Unpublished Working Paper.*

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

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

Chamberlain, Gary. 1986. "Asymptotic Efficiency in Semiparametric Models with Censoring." *Journal of Econometrics*, 32(2): 189–218.

Chan, David, Matthew Gentzkow, and Chuan Yu. 2020. "Selection with Variation in Diagnostic Skill: Evidence from Radiologists." *NBER Working Paper No. 26467.*

Dobbie, Will, Jacob Goldin, and Crystal Yang. 2018. "The Effects of Pre-Trial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges." *American Economic Review*, 108(2): 201–240.

Ewens, Michael, Bryan Tomlin, and Liang Choon Wang. 2014. "Statistical Discrimination or Prejudice? A Large Sample Field Experiment." *Review of Economics and Statistics*, 96(1): 119–134.

Feigenberg, Benjamin, and Conrad Miller. 2020. "Racial Disparities in Motor Vehicle Searches Cannot Be Justified By Efficiency." *Unpublished Working Paper.*

Frandsen, Brigham, Lars Lefgren, and Emily Leslie. 2019. "Judging Judge Fixed Effects." *NBER Working Paper No. 25528.*

Gelman, Andrew, Jeffrey Fagan, and Alex Kiss. 2007. "An Analysis of the New York City Police Department's "Stop-and-Frisk" Policy in the Context of Claims of Racial Bias." *Journal of the American Statistical Association*, 102(479): 813–823.

Heckman, James J. 1990. "Varieties of Selection Bias." *American Economic Review Papers and Proceedings*, 80(2): 313–318.

Heckman, James J., and Edward Vytlacil. 2005. "Structural Equations, Treatment Effects, and Econometric Policy Evaluation." *Econometrica*, 73(3): 669–738.

Hull, Peter. 2020. "Estimating Hospital Quality with Quasi-Experimental Data." *Unpublished Working Paper.*

Imbens, Guido, and Joshua Angrist. 1994. "A Least Squares Correction for Selectivity Bias." *Econometrica*, 62(2): 467–475.

Kitagawa, Toru. 2015. "A Test for Instrument Validity." *Econometrica*, 83(5): 2043–2063.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. "Human Decisions and Machine Predictions." *Quarterly Journal of Economics*, 133(1): 237–293.

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. 2017. "Inherent Trade-Offs in Algorithmic Fairness." *Proceedings of the 8th Conference on Innovations in Theoretical Computer Science.*

Kowalski, Amanda. 2016. "Doing More When You're Running LATE: Applying Marginal Treatment Effect Methods to Examine Treatment Effect Heterogeneity in Experiments." *NBER Working Paper No. 22363.*

Leslie, Emily, and Nolan Pope. 2017. "The Unintended Impact of Pretrial Detention on Case Outcomes: Evidence from NYC Arraignments." *Journal of Law and Economics*, 60(3): 529–557.

Marx, Philip. 2018. "An Absolute Test of Racial Prejudice." *Unpublished Working Paper.*

McIntyre, Frank, and Shima Baradaran. 2013. "Race, Prediction, and Pretrial Detention." *Journal of Empirical Legal Studies*, 10(4): 741–770.

Mogstad, Magne, Alexander Torgovitsky, and Christopher Walters. 2019. "Identification of Causal Effects with Multiple Instruments: Problems and Some Solutions." *NBER Working Paper No. 25691.*

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

Morris, Carl. 1983. "Parametric Empirical Bayes Inference: Theory and Applications." *Journal of the American Statistical Association*, 78(381): 47–55.

Mueller-Smith, Michael. 2015. "The Criminal and Labor Market Impacts of Incarceration." *Unpublished Working Paper.*

New York City Criminal Justice Agency Inc. 2016. "Annual Report 2015."

Norris, Sam. 2019. "Examiner Inconsistency: Evidence from Refugee Appeals." *Unpublished Working Paper.*

Pakes, Ariel, and David Pollard. 1989. "Simulation and the Asymptotics of Optimization Estimators." *Econometrica*, 57(5): 1027–1057.

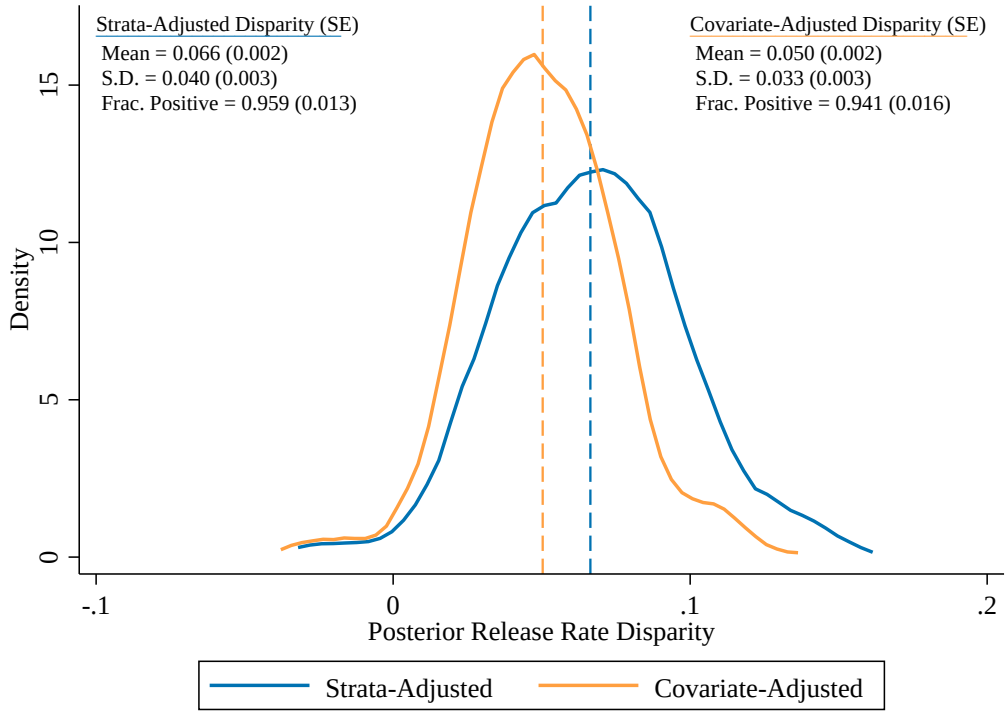
Phelps, Edmund S. 1972. "The Statistical Theory of Racism and Sexism." *American Economic Review*, 62(4): 659–661.

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

Rose, Evan. 2020. "Who Gets a Second Chance? Effectiveness and Equity in Supervision of Criminal Offenders." *Unpublished Working Paper.*

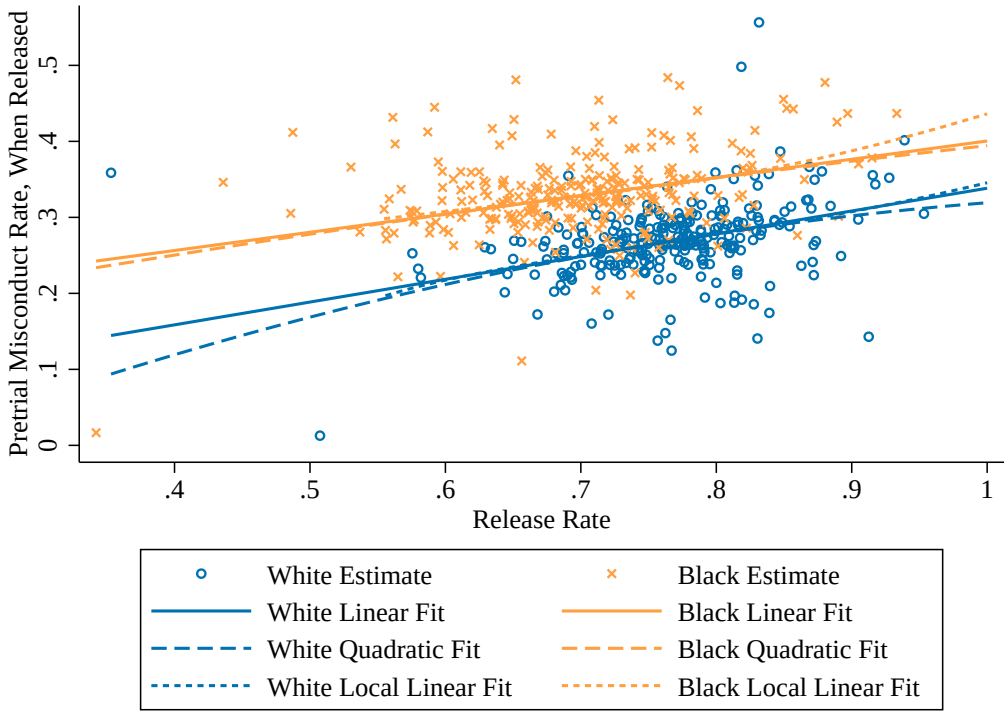
Yang, Crystal, and Will Dobbie. 2019. "Equal Protection Under Algorithms: A New Statistical and Legal Framework." *Unpublished Working Paper.*

Figure 1: Observational Release Rate Disparities



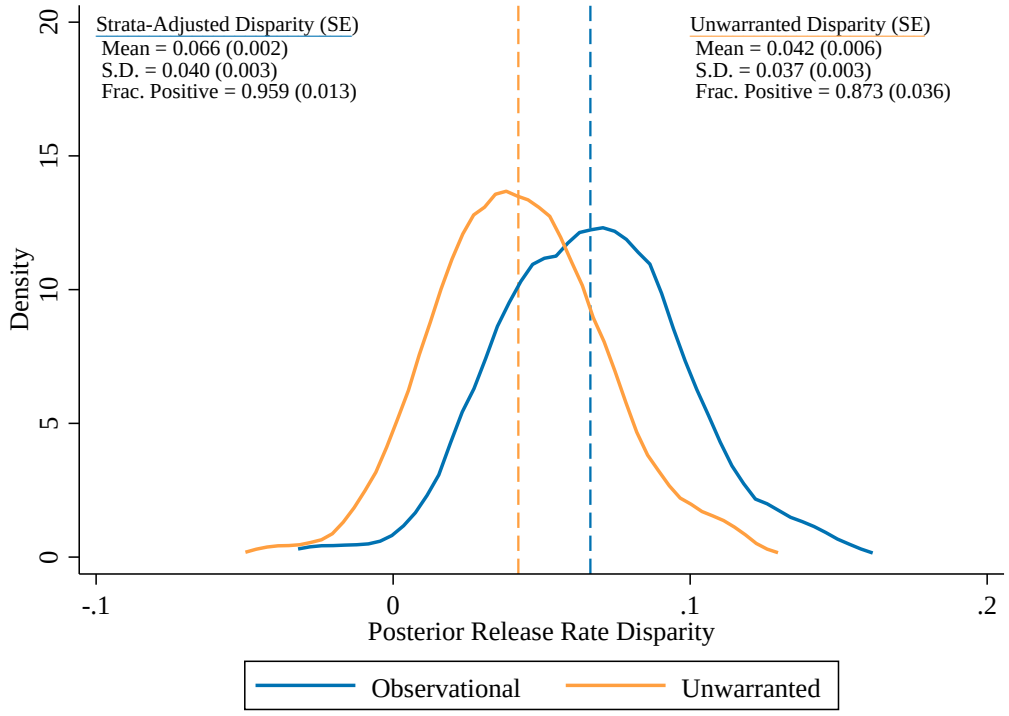
Notes. This figure plots the distribution of observational release rate disparity posteriors for the 268 judges in our sample. Estimates are from the coefficients of an OLS regression of an indicator for pretrial release on white×judge fixed effects, controlling for judge main effects. Empirical Bayes posteriors are computed using a standard shrinkage procedure, as described in Appendix B.4. The strata-adjusted disparity line shows the distribution of posteriors when controlling only for the main judge fixed effects and court-by-time fixed effects. The covariate-adjusted posterior distribution adds the baseline controls from Table 2. Means and standard deviations refer to the estimated prior distribution. The fractions of positive disparities are computed as posterior average effects, as described in Appendix B.4.

Figure 2: Judge-Specific Release Rates and Conditional Misconduct Rates



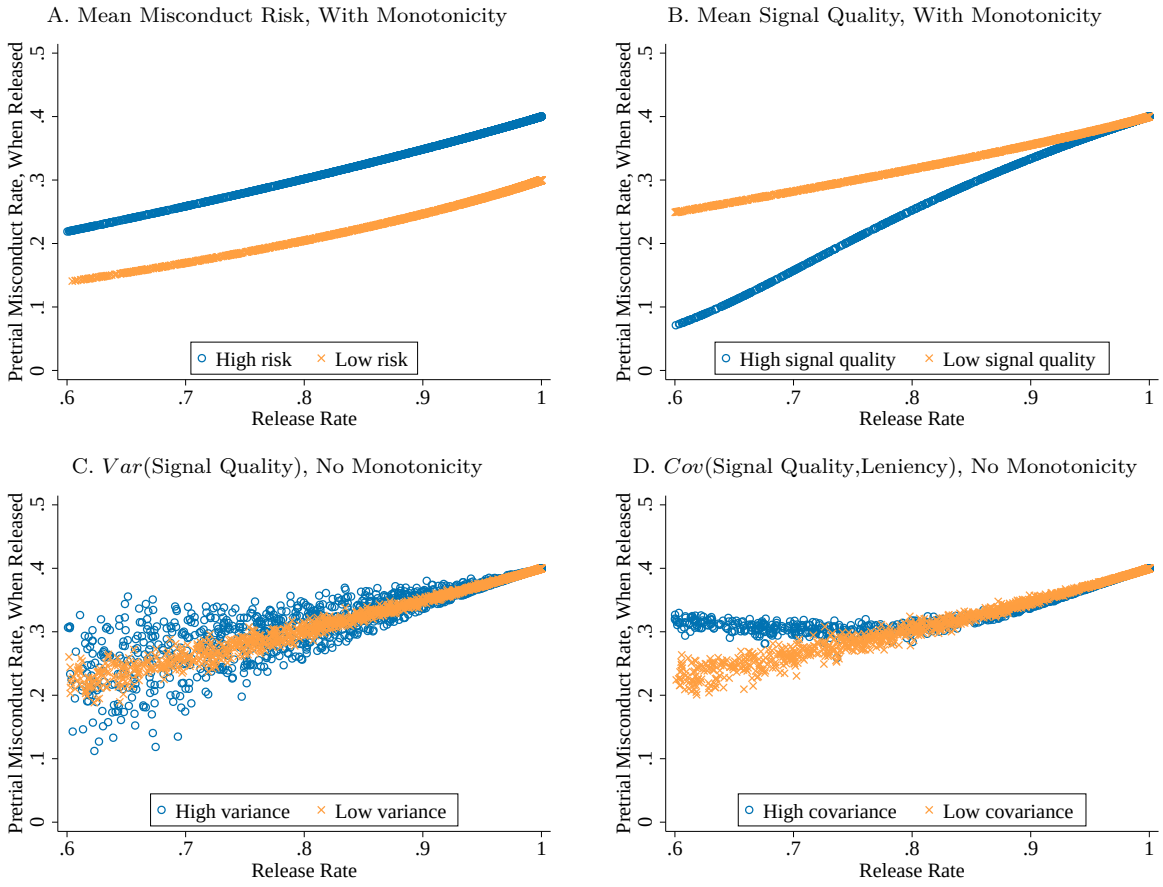
Notes. This figure plots race-specific release rates for the 268 judges in our sample against rates of pretrial misconduct among the set of released defendants. All estimates adjust for court-by-time fixed effects. The figure also plots race-specific linear, quadratic, and local linear curves of best fit, obtained from judge-level regressions that inverse-weight by the variance of the estimated misconduct rate among released defendants. The local linear regressions use a Gaussian kernel with a race-specific rule-of-thumb bandwidth.

Figure 3: Observational and Unwarranted Release Rate Disparities



Notes. This figure plots the distribution of observational and unwarranted release rate disparity posteriors for the 268 judges in our sample. Strata-adjusted disparities are estimated by the coefficients of an OLS regression of an indicator for pretrial release on white×judge fixed effects, controlling for judge main effects and court-by-time fixed effects. Unwarranted disparities are estimated as described in Section 5, using the local linear extrapolations from Figure 2 to estimate the mean risk of each race. Empirical Bayes posteriors are computed using a standard shrinkage procedure, as described in Appendix B.4. Means and standard deviations refer to the estimated prior distribution. The fractions of positive disparities are computed as posterior average effects, as described in Appendix B.4.

Figure 4: Identification of Hierarchical MTE Model Parameters



Notes. This figure plots simulated race- and judge-specific release rates against rates of pretrial misconduct among the set of released defendants under different parameterizations of the hierarchical MTE model described in the text. Panel A plots differences in mean misconduct risk ($\mu = 0.4$ vs. $\mu = 0.3$) when conventional MTE monotonicity holds ($\psi = 0$). Panel B plots differences in mean signal quality ($\alpha = 1$ vs. $\alpha = 0$) when conventional MTE monotonicity holds ($\psi = 0$). Panel C plots differences in signal quality variance ($\psi = 0.4$ vs. $\psi = 0.1$). Panel D plots differences in the covariance between judge signal quality and judge leniency ($\beta = 2$ vs. $\beta = 0.1$). The default parameterization is $\mu = 0.4$, $\alpha = 0.2$, $\psi = 0.1$, $\beta = 0$, $\gamma = 1.3$, and $\delta = 1$.

Table 1: Descriptive Statistics

	All Defendants	White Defendants	Black Defendants
Panel A: Pretrial Release	(1)	(2)	(3)
Released Before Trial	0.730	0.767	0.695
Share ROR	0.852	0.852	0.851
Share Money Bail	0.144	0.144	0.145
Share Other Bail Type	0.004	0.004	0.004
Share Remanded	0.000	0.000	0.000
Panel B: Defendant Characteristics			
White	0.478	1.000	0.000
Male	0.821	0.839	0.804
Age at Arrest	31.97	32.06	31.89
Prior Rearrest	0.229	0.204	0.253
Prior FTA	0.103	0.087	0.117
Panel C: Charge Characteristics			
Number of Charges	1.150	1.184	1.118
Felony Charge	0.362	0.355	0.368
Misdemeanor Charge	0.638	0.645	0.632
Any Drug Charge	0.256	0.257	0.256
Any DUI Charge	0.046	0.067	0.027
Any Violent Charge	0.143	0.124	0.160
Any Property Charge	0.136	0.127	0.144
Panel D: Pretrial Misconduct, When Released			
Pretrial Misconduct	0.299	0.266	0.332
Share Rearrest Only	0.499	0.498	0.499
Share FTA Only	0.281	0.296	0.269
Share Rearrest and FTA	0.220	0.205	0.232
Total Cases	595,186	284,598	310,588
Cases with Defendant Released	434,201	218,256	215,945

Notes. This table summarizes the NYC analysis sample. The sample consists of bail hearings that were quasi-randomly assigned judges between November 1, 2008 and November 1, 2013, as described in the text. Information on demographics and criminal outcomes is derived from court records as described in the text. Pretrial release is defined as meeting the bail conditions set by the first assigned bail judge. ROR (released on recognizance) is defined as being released without any conditions. FTA (failure to appear) is defined as failing to appear at a mandated court date.

Table 2: Observational Release Rate Disparities

	(1)	(2)	(3)
White	0.072	0.068	0.052
	(0.005)	(0.005)	(0.004)
Male			-0.092
			(0.004)
Age at Arrest			-0.005
			(0.000)
Prior Rearrest			-0.068
			(0.004)
Prior FTA			-0.208
			(0.005)
Felony Charge			-0.171
			(0.005)
Any Drug Charge			-0.057
			(0.007)
Any DUI Charge			0.119
			(0.004)
Any Violent Charge			-0.146
			(0.007)
Any Property Charge			-0.072
			(0.005)
Court x Time FE	No	Yes	Yes
Case/Defendant Observables	No	No	Yes
Mean Release Rate	0.730	0.730	0.730
Cases	595,186	595,186	595,186

Notes. This table reports OLS estimates of regressions of an indicator for pretrial release on defendant characteristics. The regressions are estimated on the sample described in the notes to Table 1. Robust standard errors, two-way clustered at the individual and the judge level, are reported in parentheses.

Table 3: Mean Risk and Unwarranted Disparity Estimates

	Linear Extrapolation	Quadratic Extrapolation	Local Linear Extrapolation
Panel A: Mean Risk by Race	(1)	(2)	(3)
White Defendants	0.338	0.319	0.346
	(0.007)	(0.021)	(0.014)
Black Defendants	0.400	0.394	0.436
	(0.006)	(0.021)	(0.016)
Panel B: System-Wide Discrimination			
Mean Across Cases	0.054	0.054	0.042
	(0.002)	(0.007)	(0.006)
Panel C: Judge-Level Discrimination			
Mean Across Judges	0.054	0.054	0.042
	(0.003)	(0.007)	(0.006)
Std. Dev. Across Judges	0.038	0.037	0.037
	(0.003)	(0.003)	(0.003)
Fraction Positive	0.929	0.931	0.873
	(0.016)	(0.036)	(0.036)
Judges	268	268	268

Notes. This table summarizes estimates of mean risk and unwarranted racial disparities from different extrapolations of the variation in Figure 2. Panel A reports estimates of race-specific average misconduct risk, Panel B reports estimates of system-wide (case-weighted) unwarranted disparity, and Panel C reports empirical Bayes estimates of summary statistics for the judge-level unwarranted disparity prior distribution. To estimate mean risk, column 1 uses a linear extrapolation of the variation in Figure 2, while column 2 uses a quadratic extrapolation and column 3 uses a local linear extrapolation with a Gaussian kernel and a rule-of-thumb bandwidth. Robust standard errors, two-way clustered at the individual and judge level, are obtained by a bootstrapping procedure and appear in parentheses.

Table 4: Unwarranted Disparities and Judge Characteristics

	Full-Sample Disparities					Split-Sample Disparities	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
New Judge	-0.016				-0.012		-0.006
	(0.005)				(0.004)		(0.005)
Lenient Judge		-0.009			-0.011		-0.003
		(0.004)			(0.003)		(0.003)
Above-Median Black Share			-0.012		-0.007		0.003
			(0.004)		(0.004)		(0.004)
Manhattan Courtroom				0.036	0.033		0.022
				(0.005)	(0.004)		(0.005)
Bronx Courtroom				-0.004	-0.007		0.006
				(0.004)	(0.005)		(0.006)
Queens Courtroom				0.028	0.022		0.022
				(0.005)	(0.006)		(0.005)
Richmond Courtroom				0.016	0.010		0.022
				(0.004)	(0.007)		(0.006)
Lagged Disparity						0.556	0.380
						(0.062)	(0.069)
Mean Disparity	0.042	0.042	0.042	0.042	0.042	0.056	0.056
R2	0.048	0.023	0.041	0.348	0.414	0.251	0.348

Notes. This table reports OLS estimates of regressions of unwarranted disparity posteriors on judge characteristics. Unwarranted disparities are estimated as described in Section 5, using the benchmark local linear estimate of mean risk. New judges are defined as judges appointed during our estimation period. Lenient judges are defined as judges with above-average leniency, controlling for court-by-time fixed effects. Courtroom locations are defined using the location of the modal case heard by each judge. Split-sample disparities are computed by splitting each judge's sample of cases at the median case and constructing two samples, a before-median case sample and an after-median case sample. Unwarranted disparities are then re-estimated within each subsample. The estimation procedure conditions on court-by-time effects, which causes a small number of judge effects to become collinear with the court-by-time effects and dropped. All specifications are weighted by the inverse variance of the unwarranted disparity posteriors. Empirical Bayes posteriors are computed using a standard shrinkage procedure (Morris, 1983). Robust standard errors are reported in parentheses.

Table 5: Hierarchical MTE Model Estimates

	With Monotonicity			Without Monotonicity		
	White Defendants	Black Defendants	Diff.	White Defendants	Black Defendants	Diff.
	(1)	(2)	(3)	(4)	(5)	(6)
Mean Misconduct Risk	0.346	0.423	-0.077	0.391	0.441	-0.050
	(0.008)	(0.009)	(0.012)	(0.007)	(0.007)	(0.010)
Mean Marginal Released Outcome	0.616	0.511	0.105	0.651	0.576	0.074
	(0.057)	(0.030)	(0.061)	(0.033)	(0.021)	(0.038)
Mean Signal Quality	1.712	0.963	0.749	1.385	0.970	0.416
	(0.219)	(0.141)	(0.271)	(0.104)	(0.073)	(0.128)
Marginal Outcome Std. Dev.	0.211	0.094	0.117	0.080	0.064	0.016
	(0.029)	(0.022)	(0.037)	(0.009)	(0.005)	(0.010)
Signal Quality Std. Dev.				0.196	0.163	0.033
				(0.038)	(0.017)	(0.041)
Covariance of Signal Quality and Marginal Released Outcomes				0.013	0.007	0.006
				(0.005)	(0.002)	(0.005)
Judges	268	268	–	268	268	–

Notes. This table reports simulated minimum distance estimates of moments of the MTE model described in Section 6. See Table A16 for underlying hyperparameter estimates. Columns 4-6 estimate the baseline model, while columns 1-3 impose conventional monotonicity. Robust standard errors, two-way clustered at the individual and the judge level, are obtained by a bootstrapping procedure and appear in parentheses.

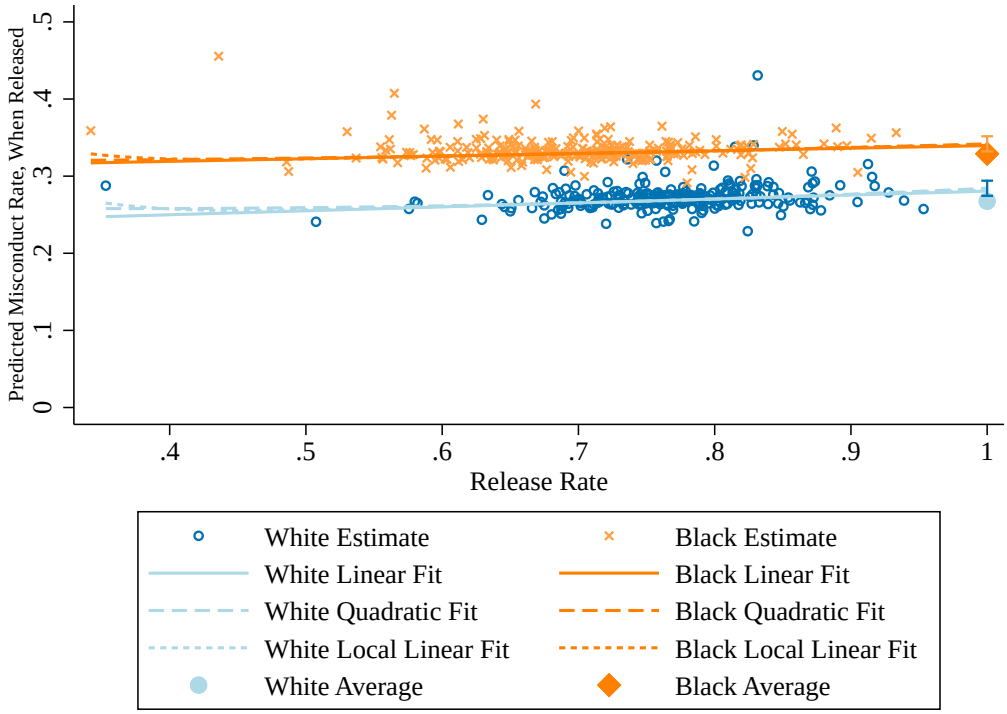
Table 6: Policy Simulations

	Baseline	Target Unwarranted Disparity Posteriors		Target Observational Disparity Posteriors	
		Increase Leniency	Decrease Leniency	Increase Leniency	Decrease Leniency
Panel A: Close All Disparities	(1)	(2)	(3)	(4)	(5)
Mean Unwarranted Disparity	0.047	0.000	0.000	-0.017	-0.019
	[0.037]	[0.020]	[0.026]	[0.020]	[0.026]
Mean Observational Disparity	0.065	0.017	0.019	0.000	-0.000
	[0.038]	[0.020]	[0.026]	[0.019]	[0.026]
Racial Bias	0.074	0.039	0.013	0.025	-0.011
	[0.078]	[0.068]	[0.055]	[0.070]	[0.053]
Panel B: Close Top-Quintile Disparities					
Mean Unwarranted Disparity		0.030	0.030	0.026	0.026
		[0.035]	[0.037]	[0.038]	[0.041]
Mean Observational Disparity		0.047	0.048	0.044	0.043
		[0.035]	[0.037]	[0.039]	[0.040]
Racial Bias		0.062	0.051	0.059	0.045
		[0.075]	[0.076]	[0.076]	[0.080]
Observations	268	268	268	268	268

Notes. This table reports the results from a series of policy simulations. Column 1 reports the mean unwarranted disparity, observational disparity, and racial bias across judges and 250 simulations of the hierarchical MTE model. Average standard deviations across judges are included in brackets. Simulations are based on the estimates from columns 2 and 4 of Appendix Table A16. Column 2 of Panel A recomputes the statistics for a counterfactual in which the lower of the Black or white release rate of each judge is raised to equalize unwarranted disparity posteriors, while column 3 of Panel A does the same by lowering one of the two release rates. Columns 4 and 5 of Panel A instead adjust release rates to equalize observational disparity posteriors. Panel B conducts the counterfactual exercises only on judges ranked in the top quintile of unwarranted (columns 2 and 3) or observational (columns 4 and 5) disparity posteriors. Estimates of the model hyperparameters and empirical Bayes posteriors of all judge-specific parameters are recomputed in each simulation draw via the SMD procedure outlined in the text, using moments simulated according to the estimated distribution of reduced-form estimates in Figure 2.

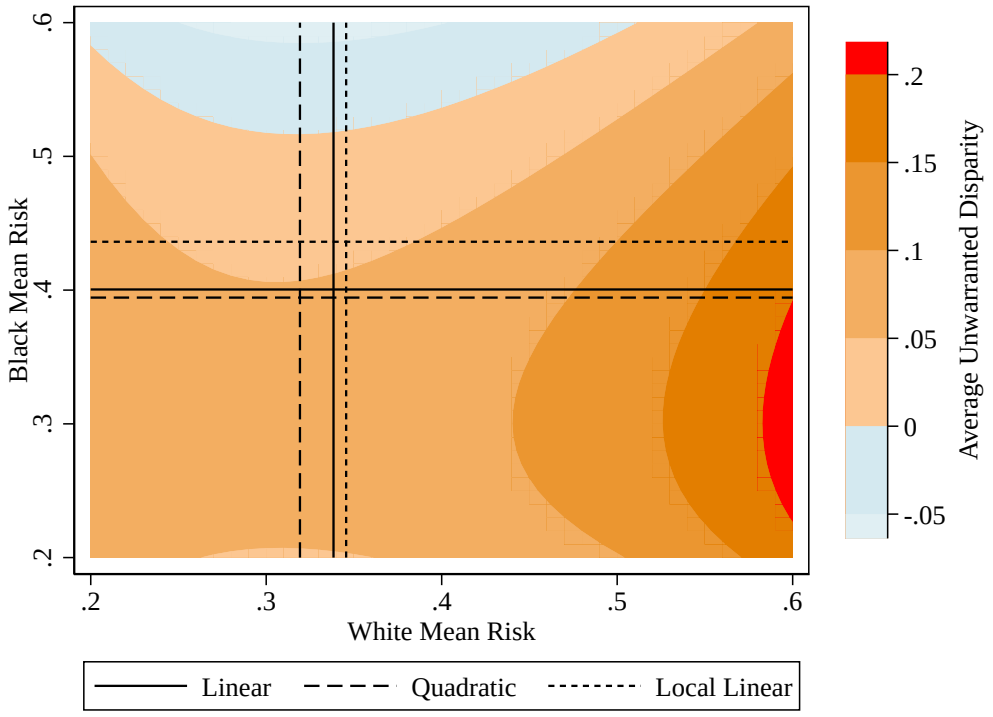
A Appendix Figures and Tables

Appendix Figure A1: Placebo Mean Risk Extrapolation



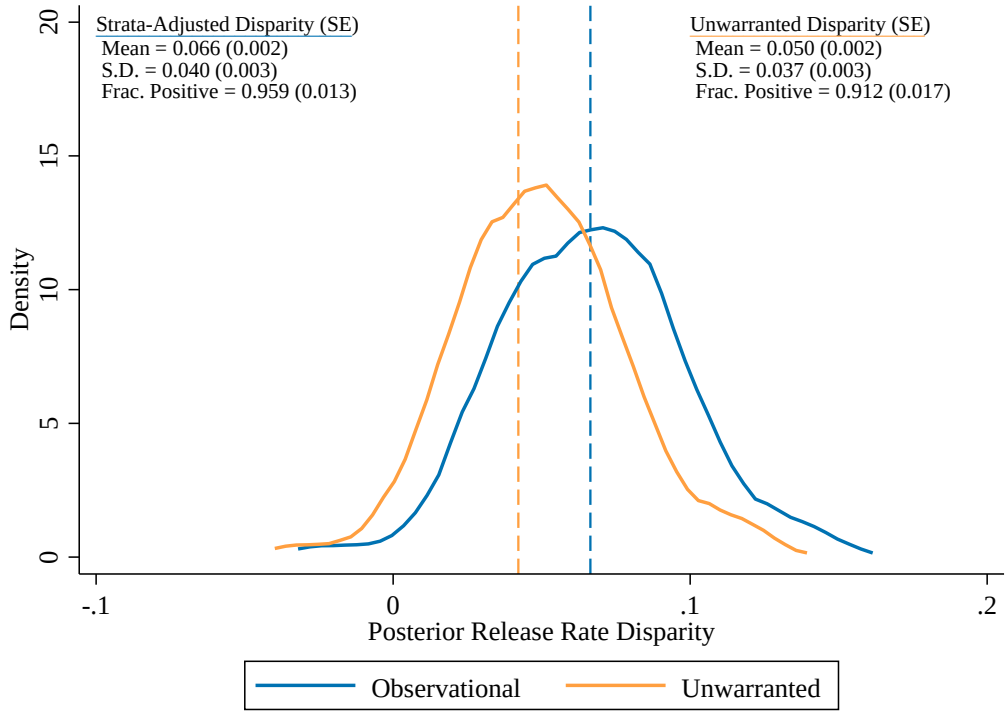
Notes. This figure plots race-specific release rates for the 268 judges in our sample against rates of predicted pretrial misconduct among the set of released defendants. Predicted misconduct is given by the fitted values of an OLS regression of misconduct on the regressors in column 3 of Table 2, estimated in the set of released defendants. Average predicted misconduct rates in the full sample of white and Black defendants are indicated with solid markers at the maximal release rate of one. All estimates adjust for court-by-time fixed effects. The figure also plots race-specific linear, quadratic, and local linear curves of best fit, obtained from judge-level regressions that inverse-weight by the variance of the estimated predicted misconduct rate among released defendants. The local linear regression uses a Gaussian kernel with a race-specific rule-of-thumb bandwidth. 95 percent confidence intervals for the local linear extrapolations' intercept estimates at one, obtained from robust standard errors two-way clustered at the individual and judge level, are indicated with brackets.

Appendix Figure A2: Sensitivity Analysis



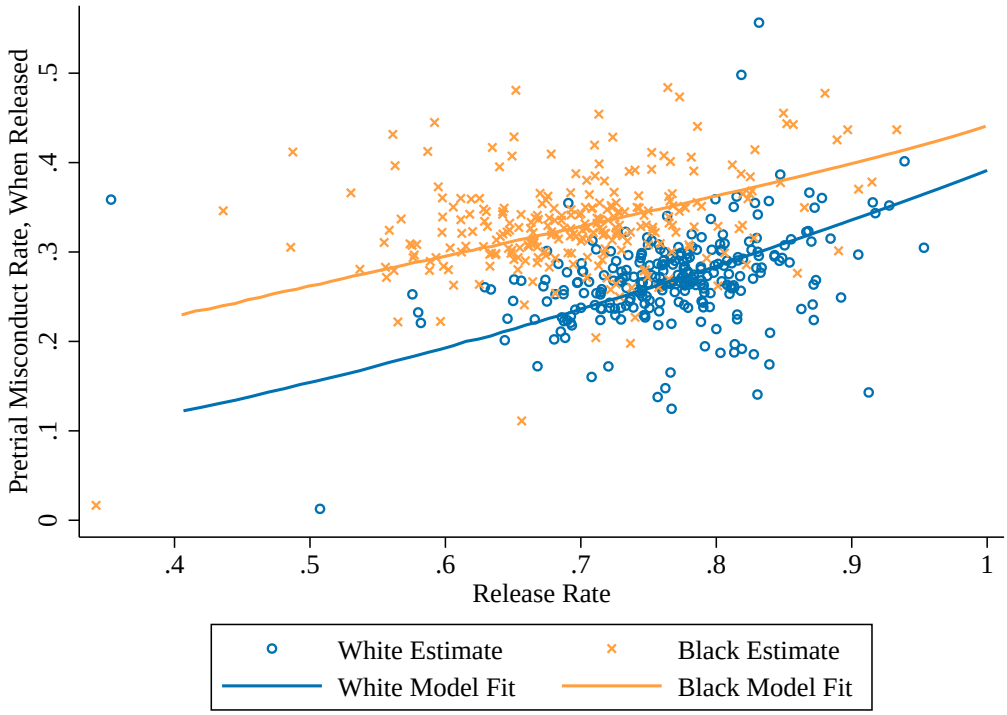
Notes. This figure shows how our estimate of system-wide discrimination changes under different estimates of white and Black mean risk. The mean risk estimates obtained from the linear, quadratic, and local linear extrapolations in Figure 2 are indicated by solid, dashed, and dotted lines.

Appendix Figure A3: Unwarranted Release Rate Disparities, Model-Based Mean Risk Estimates



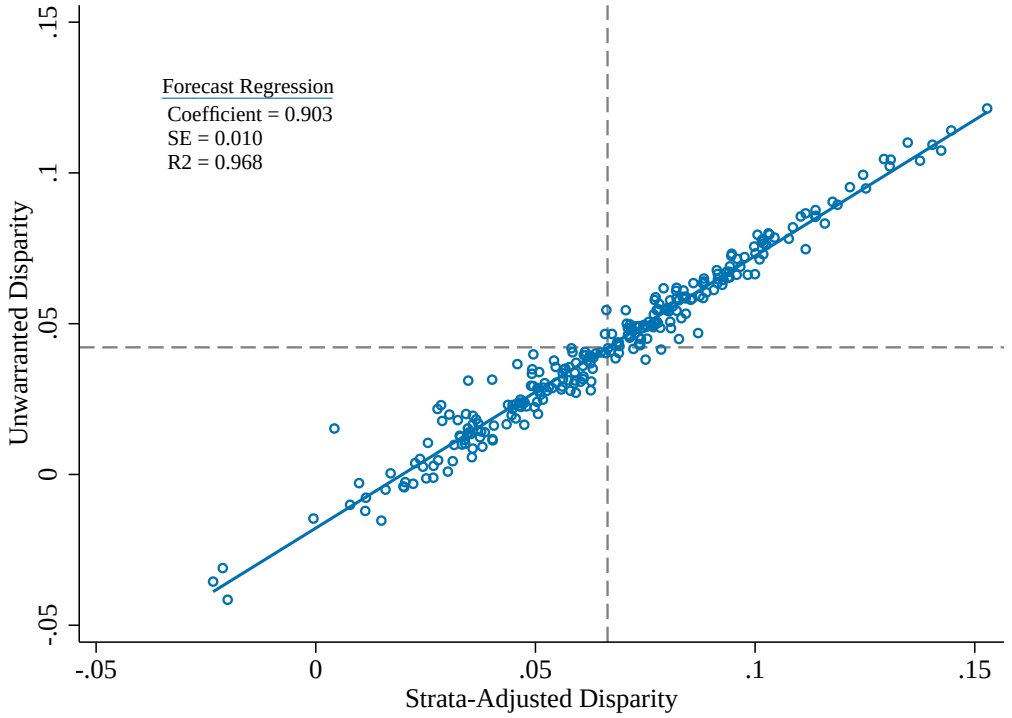
Notes. This figure plots the distribution of observational and unwarranted release rate disparity posteriors for the 268 judges in our sample. Strata-adjusted disparities are estimated by the coefficients of an OLS regression of an indicator for pretrial release on white×judge fixed effects, controlling for judge main effects and court-by-time fixed effects. Unwarranted disparities are estimated as described in Section 5, using the hierarchical MTE model estimates of mean risk for each race. Empirical Bayes posteriors are computed using a standard shrinkage procedure, as described in Appendix B.4. Means and standard deviations refer to the estimated prior distribution. The fractions of positive disparities are computed as posterior average effects, as described in Appendix B.4

Appendix Figure A4: Hierarchical MTE Model Fit



Notes. This figure plots race-specific release rates for the 268 judges in our sample against rates of pretrial misconduct among the set of released defendants. All estimates adjust for court-by-time fixed effects. The figure also plots race-specific curves of best fit implied by our baseline hierarchical MTE model hyperparameter estimates.

Appendix Figure A5: Predictiveness of Observational Release Rate Disparities



Notes. This figure plots unwarranted white-black release rate disparity posteriors against the corresponding strata-adjusted release rate disparity posteriors for the 268 judges in our sample. Observational disparities are estimated by the coefficients of an OLS regression of an indicator for pretrial release on white×judge fixed effects, controlling for judge main effects and court-by-time fixed effects. Unwarranted disparities are estimated as described in Section 5, using the local linear extrapolation from Figure 2 to estimate the mean risk of each race. Empirical Bayes posteriors are computed using a standard shrinkage procedure, as described in Appendix B.4. The slope of the solid line indicates the forecast coefficient.

Appendix Table A1: Judge Leniency and Sample Attrition

	All Defendants	White Defendants	Black Defendants
	(1)	(2)	(3)
Dropped from Sample	0.00007	0.00003	0.00012
	(0.00012)	(0.00013)	(0.00014)
Court x Time FE	Yes	Yes	Yes
Mean Sample Attrition	0.416	0.409	0.424
Cases	1,425,652	726,284	697,597

Notes. This table reports OLS estimates of regressions of judge leniency on an indicator for leaving the sample due to case adjournment or case disposal and court-by-time fixed effects. The regressions are estimated on the sample of all arraignments made in NYC between November 1, 2008 and November 1, 2013. Judge leniency is estimated using data from other cases assigned to a given bail judge, following the procedure described in Section 4.1. Robust standard errors, two-way clustered at the individual and the judge level, are reported in parentheses.

Appendix Table A2: Descriptive Statistics by Sample

	All Defendants		White Defendants		Black Defendants	
	Full Sample	Estimation Sample	Full Sample	Estimation Sample	Full Sample	Estimation Sample
Panel A: Pretrial Release	(1)	(2)	(3)	(4)	(5)	(6)
Released Before Trial	0.852	0.730	0.872	0.767	0.832	0.695
Share ROR	0.601	0.852	0.616	0.852	0.586	0.851
Share Disposed	0.301	0.000	0.274	0.000	0.327	0.000
Share Adjourned	0.191	0.000	0.199	0.000	0.183	0.000
Share Money Bail	0.068	0.144	0.070	0.144	0.066	0.145
Share Other Bail Type	0.332	0.004	0.314	0.004	0.348	0.004
Share Remanded	0.000	0.000	0.000	0.000	0.000	0.000
Panel B: Defendant Characteristics						
White	0.483	0.478	1.000	1.000	0.000	0.000
Male	0.822	0.821	0.831	0.839	0.813	0.804
Age at Arrest	31.819	31.969	31.540	32.055	32.080	31.890
Prior Rearrest	0.192	0.229	0.168	0.204	0.214	0.253
Prior FTA	0.085	0.103	0.071	0.087	0.099	0.117
Panel C: Charge Characteristics						
Number of Charges	1.094	1.150	1.111	1.184	1.078	1.118
Felony Charge	0.184	0.362	0.181	0.355	0.188	0.368
Misdemeanor Charge	0.816	0.638	0.819	0.645	0.812	0.632
Any Drug Charge	0.347	0.256	0.342	0.257	0.352	0.256
Any DUI Charge	0.031	0.046	0.046	0.067	0.017	0.027
Any Violent Charge	0.072	0.143	0.062	0.124	0.081	0.160
Any Property Charge	0.217	0.136	0.209	0.127	0.226	0.144
Cases	1,358,278	595,186	656,711	284,598	701,567	310,588

Notes. This table summarizes the difference between the NYC analysis sample and the full sample of NYC arraignments. The full sample consists of all bail hearings between November 1, 2008 and November 1, 2013. The analysis sample consists of bail hearings that were quasi-randomly assigned to judges between November 1, 2008 and November 1, 2013, as described in the text. Information on demographics and criminal outcomes is derived from court records as described in the text. Pretrial release is defined as meeting the bail conditions set by the first assigned bail judge. ROR (released on Recognizance) is defined as being released without any conditions. FTA (failure to appear) is defined as failing to appear at a mandated court date.

Appendix Table A3: Tests of Quasi-Random Judge Assignment

	All Defendants	White Defendants	Black Defendants
	(1)	(2)	(3)
White	0.00013		
	(0.00009)		
Male	0.00003	0.00003	0.00004
	(0.00014)	(0.00019)	(0.00018)
Age at Arrest	-0.00011	-0.00015	-0.00008
	(0.00004)	(0.00006)	(0.00005)
Prior Rearrest	-0.00021	0.00006	-0.00042
	(0.00011)	(0.00018)	(0.00015)
Prior FTA	0.00016	-0.00011	0.00036
	(0.00016)	(0.00024)	(0.00023)
Number of Charges	-0.00001	-0.00001	-0.00001
	(0.00001)	(0.00001)	(0.00003)
Felony Charge	0.00025	0.00011	0.00039
	(0.00020)	(0.00023)	(0.00025)
Any Drug Charge	-0.00022	-0.00017	-0.00027
	(0.00016)	(0.00021)	(0.00018)
Any DUI Charge	0.00045	0.00051	0.00008
	(0.00027)	(0.00032)	(0.00045)
Any Violent Charge	-0.00008	-0.00023	0.00001
	(0.00023)	(0.00033)	(0.00025)
Any Property Charge	-0.00033	-0.00028	-0.00036
	(0.00018)	(0.00019)	(0.00027)
Joint p-value	[0.10689]	[0.29792]	[0.10136]
Court x Time FE	Yes	Yes	Yes
Cases	595,186	284,598	310,588

Notes. This table reports OLS estimates of regressions of judge leniency on defendant characteristics. The regressions are estimated on the sample described in Table 1. Judge leniency is estimated using data from other cases assigned to a given bail judge, following the procedure described in Section 4.1. All regressions control for court-by-time fixed effects. The p-values reported at the bottom of each column are from F-tests of the joint significance of the variables listed in the rows. Robust standard errors, two-way clustered at the individual and the judge level, are reported in parentheses.

Appendix Table A4: First Stage Effects of Judge Leniency

	All Defendants	White Defendants	Black Defendants
	(1)	(2)	(3)
Judge Leniency	0.960	0.788	1.104
	(0.025)	(0.029)	(0.033)
Court x Time FE	Yes	Yes	Yes
Mean Release Rate	0.730	0.767	0.695
Cases	595,186	284,598	310,588

Notes. This table reports OLS estimates of regressions of an indicator for pretrial release on judge leniency. The regressions are estimated on the sample described in Table 1. Judge leniency is estimated using data from other cases assigned to a bail judge, following the procedure described in Section 4.1. All regressions control for court-by-time fixed effects. Robust standard errors, two-way clustered at the individual and the judge level, are reported in parentheses.

Appendix Table A5: Simple Numerical Example of Unwarranted Disparity Estimation

		Number of Defendants	Number Released	Scaling Factor	Rescaled Released	Release Rate	Release Disparity
Panel A: Observational Estimates		(1)	(2)	(3)	(4)	(5)	(6)
Low-Risk Defendants	$Y_i^* = 0$	75	60	1	60	0.65	0.30
	$Y_i^* = 1$	25	5	1	5		
High-Risk Defendants	$Y_i^* = 0$	25	20	1	20	0.35	
	$Y_i^* = 1$	75	15	1	15		
Panel B: Rescaled Estimates							
Low-Risk Defendants	$Y_i^* = 0$	75	60	2/3	40	0.50	0.00
	$Y_i^* = 1$	25	5	2	10		
High-Risk Defendants	$Y_i^* = 0$	25	20	2	40	0.50	
	$Y_i^* = 1$	75	15	2/3	10		

Notes: This table uses a simple numerical example to illustrate how unwarranted disparities can be measured with observational release rate comparisons that are rescaled using average group-specific misconduct risk. We assume there is one type-neutral judge who releases 80 percent of defendants with $Y_i^* = 0$ and 20 percent of defendants with $Y_i^* = 1$. The judge observes the type of the defendant, which is either High-risk or Low-risk. There are 100 High-risk defendants where 75 have $Y_i^* = 1$, and 100 Low-risk defendants where 25 have $Y_i^* = 1$. Panel A shows that the judge has a Low-risk release rate of 0.65 but a High-risk release rate of 0.35, meaning that an observational comparison would find that Low-risk defendants have a 30 percentage point higher release rate than High-risk defendants despite the judge being type-neutral. Panel B shows that the true unwarranted disparity of zero can be measured by rescaling this observational release rate comparison with the scaling factor described in the text. Column 3 of Panel B shows the scaling factor (Ω_i) in this example, and column 6 shows the resulting unwarranted disparity estimate.

Appendix Table A6: Mean Risk and Unwarranted Disparity Estimates, Borough-Specific Estimates

	Linear Extrapolation	Quadratic Extrapolation	Local Linear Extrapolation
Panel A: Mean Risk by Race	(1)	(2)	(3)
White Defendants	0.337	0.342	0.337
	(0.014)	(0.034)	(0.025)
Black Defendants	0.414	0.400	0.419
	(0.009)	(0.023)	(0.021)
Panel B: System-Wide Discrimination			
Mean Across Cases	0.050	0.052	0.046
	(0.008)	(0.020)	(0.010)
Panel C: Judge-Level Discrimination			
Mean Across Judges	0.043	0.048	0.041
	(0.032)	(0.072)	(0.028)
Std. Dev. Across Judges	0.033	0.040	0.040
	(0.032)	(0.072)	(0.028)
Fraction Positive	0.904	0.883	0.848
	(0.025)	(0.050)	(0.049)
Judges	268	268	268

Notes. This table summarizes estimates of mean risk and unwarranted racial disparities. This tables estimates conditional regression models for each borough and averages the resulting estimates by borough share. Panel A reports estimates of race-specific average misconduct risk, Panel B reports estimates of system-wide (case-weighted) unwarranted disparity, and Panel C reports empirical Bayes estimates of summary statistics for the judge-level unwarranted disparity prior distribution. To estimate mean risk, column 1 uses a linear extrapolation of the variation in Figure 2, while column 2 uses a quadratic extrapolation and column 3 uses a local linear extrapolation with a Gaussian kernel and a rule-of-thumb bandwidth. Robust standard errors, two-way clustered at the individual and judge level, are obtained by a bootstrapping procedure and appear in parentheses.

Appendix Table A7: Mean Risk and Unwarranted Disparity Estimates, Shrunk Leniency Estimates

	Linear Extrapolation	Quadratic Extrapolation	Local Linear Extrapolation
Panel A: Mean Risk by Race	(1)	(2)	(3)
White Defendants	0.342	0.368	0.358
	(0.008)	(0.037)	(0.013)
Black Defendants	0.403	0.436	0.441
	(0.007)	(0.028)	(0.015)
Panel B: System-Wide Discrimination			
Mean Across Cases	0.054	0.046	0.042
	(0.003)	(0.014)	(0.006)
Panel C: Judge-Level Discrimination			
Mean Across Judges	0.053	0.046	0.042
	(0.003)	(0.014)	(0.006)
Std. Dev. Across Judges	0.029	0.029	0.029
	(0.002)	(0.002)	(0.002)
Fraction Positive	0.963	0.938	0.920
	(0.011)	(0.085)	(0.040)
Judges	268	268	268

Notes. This table summarizes estimates of mean risk and unwarranted racial disparities from different extrapolations of the variation in Figure 2, after applying conventional empirical Bayes shrinkage to the judge- and race-specific leniency estimates. Panel A reports estimates of race-specific average misconduct risk, Panel B reports estimates of system-wide (case-weighted) unwarranted disparity, and Panel C reports empirical Bayes estimates of summary statistics for the judge-level unwarranted disparity prior distribution. To estimate mean risk, column 1 uses a linear extrapolation of the variation in Figure 2, while column 2 uses a quadratic extrapolation and column 3 uses a local linear extrapolation with a Gaussian kernel and a rule-of-thumb bandwidth. Robust standard errors, two-way clustered at the individual and judge level, are obtained by a bootstrapping procedure and appear in parentheses.

Appendix Table A8: Unwarranted Disparity Estimation for NYC Release Decisions

		Number of Defendants	Number Released	Scaling Factor	Rescaled Released	Release Rate	Release Disparity
Panel A: Observational Estimates		(1)	(2)	(3)	(4)	(5)	(6)
White Defendants	$Y_i^* = 0$	186,250	159,296	1.000	159,296	0.765	0.068
	$Y_i^* = 1$	98,348	58,425	1.000	58,425		
Black Defendants	$Y_i^* = 0$	175,120	145,528	1.000	145,528	0.697	
	$Y_i^* = 1$	135,468	70,952	1.000	70,952		
Panel B: Rescaled Estimates							
White Defendants	$Y_i^* = 0$	186,250	159,296	0.928	147,788	0.753	0.042
	$Y_i^* = 1$	98,348	58,425	1.137	66,418		
Black Defendants	$Y_i^* = 0$	175,120	145,528	1.077	156,709	0.710	
	$Y_i^* = 1$	135,468	70,952	0.901	63,905		

Notes: This table calculates system-wide unwarranted disparity in NYC by rescaling observational release rate comparisons using estimates of average white and black misconduct risk. In Panel A we use the local linear estimates of mean risk in Table 3 to estimate the number of defendants with and without misconduct potential (column 1) as well as the number of such defendants that are released (column 2). In Panle A, column 6 we display the observational release rate disparity between white and Black defendants. In Panel B we use the same mean risk estimates to rescale this observational release rate comparison with the scaling factor described in the text. Column 3 of Panel B shows the scaling factor (Ω_i) given by these estimates, and column 6 shows the resulting unwarranted disparity estimate.

Appendix Table A9: Mean Risk and Unwarranted Disparity Bounds

	From 0.90 Leniency	From 0.85 Leniency	From 0.80 Leniency
Panel A: Mean Risk by Race	(1)	(2)	(3)
White Defendants	[0.277,0.377]	[0.248,0.398]	[0.221,0.421]
	(0.004,0.004)	(0.002,0.002)	(0.001,0.001)
Black Defendants	[0.349,0.449]	[0.313,0.463]	[0.280,0.480]
	(0.006,0.006)	(0.003,0.003)	(0.002,0.002)
Panel B: System-Wide Discrimination			
Mean Across Cases	[0.035,0.073]	[0.029,0.083]	[0.021,0.092]
	(0.003,0.002)	(0.002,0.002)	(0.002,0.001)
Panel C: Judge-Level Discrimination			
Mean Across Judges	[0.035,0.073]	[0.029,0.083]	[0.021,0.091]
	(0.003,0.002)	(0.003,0.002)	(0.002,0.002)
Std. Dev. Across Judges	[0.037,0.039]	[0.037,0.042]	[0.036,0.046]
	(0.003,0.004)	(0.003,0.004)	(0.003,0.005)
Fraction Positive	[0.821,0.975]	[0.770,0.982]	[0.694,0.989]
	(0.022,0.012)	(0.019,0.010)	(0.017,0.009)
Judges	268	268	268

Notes. This table summarizes bounds on mean risk and unwarranted racial disparities estimated from the variation in Figure 2. Panel A reports bounds on race-specific average misconduct risk, Panel B reports bounds on system-wide (case-weighted) unwarranted disparity, and Panel C reports bounds on empirical Bayes estimates of summary statistics for the judge-level unwarranted disparity prior distribution. To estimate bounds on mean risk, column 1 uses a local linear fit of released misconduct rates among judges releasing 90% of white and Black defendants. Columns 2 and 3 form bounds from judges releasing 85% and 80% of white and Black defendants, respectively. The local linear regressions use a Gaussian kernel and a rule-of-thumb bandwidth. Bounds are formed under the assumption that either none or all of the detained defendants in each column have pretrial misconduct potential. Panels B and C search within these bounds to find the combination of white and Black mean risk that minimize or maximize each unwarranted disparity statistic. Robust standard errors on the endpoints of each set of bounds, two-way clustered at the individual and judge level, are obtained by a bootstrapping procedure and appear in parentheses.

Appendix Table A10: Mean Risk and Unwarranted Disparity Estimates, With Covariate Adjustment

	Linear Extrapolation	Quadratic Extrapolation	Local Linear Extrapolation
Panel A: Mean Risk by Race	(1)	(2)	(3)
White Defendants	0.351	0.334	0.352
	(0.007)	(0.021)	(0.015)
Black Defendants	0.394	0.412	0.423
	(0.006)	(0.021)	(0.016)
Panel B: System-Wide Discrimination			
Mean Across Cases	0.043	0.037	0.035
	(0.002)	(0.006)	(0.005)
Panel C: Judge-Level Discrimination			
Mean Across Judges	0.043	0.036	0.035
	(0.002)	(0.006)	(0.005)
Std. Dev. Across Judges	0.031	0.030	0.031
	(0.003)	(0.003)	(0.003)
Fraction Positive	0.923	0.891	0.878
	(0.017)	(0.042)	(0.036)
Judges	268	268	268

Notes. This table summarizes estimates of mean risk and unwarranted racial disparities from different extrapolations of the variation in Figure 2, where release and misconduct rates adjust for both the court-by-time effects and the case and defendant observables in Table 2. Panel A reports estimates of race-specific average misconduct risk, Panel B reports estimates of system-wide (case-weighted) unwarranted disparity, and Panel C reports empirical Bayes estimates of summary statistics for the judge-level unwarranted disparity prior distribution. To estimate mean risk, column 1 uses a linear extrapolation, while column 2 uses a quadratic extrapolation and column 3 uses a local linear extrapolation with a Gaussian kernel and a rule-of-thumb bandwidth. Robust standard errors, two-way clustered at the individual and judge level, are obtained by a bootstrapping procedure and appear in parentheses.

Appendix Table A11: Mean Risk and Unwarranted Disparity Estimates, Alternative Misconduct Outcomes

	Any Misconduct	Case FTA	Any Rearrest	Violent Rearrest
Panel A: Mean Risk by Race	(1)	(2)	(3)	(4)
White Defendants	0.346	0.176	0.233	0.014
	(0.015)	(0.011)	(0.019)	(0.004)
Black Defendants	0.436	0.242	0.314	0.014
	(0.016)	(0.013)	(0.020)	(0.005)
Panel B: System-Wide Discrimination				
Mean Across Cases	0.042	0.051	0.050	0.068
	(0.006)	(0.004)	(0.005)	(0.027)
Panel C: Judge-Level Discrimination				
Mean Across Judges	0.042	0.051	0.050	0.068
	(0.006)	(0.005)	(0.005)	(0.026)
Std. Dev. Across Judges	0.037	0.039	0.039	0.045
	(0.003)	(0.003)	(0.003)	(0.014)
Fraction Positive	0.873	0.913	0.910	0.948
	(0.035)	(0.024)	(0.028)	(0.052)
Judges	268	268	268	268

Notes. This table summarizes estimates of mean risk and unwarranted racial disparities for different outcome variables. Panel A reports estimates of race-specific average misconduct risk, Panel B reports estimates of system-wide (case-weighted) unwarranted disparity, and Panel C reports empirical Bayes estimates of summary statistics for the judge-level unwarranted disparity prior distribution. Column 1 adjusts for differences by race in the mean risk of any misconduct (either rearrest or FTA). Column 2 adjusts for differences by race in the mean risk of FTA. Column 3 adjusts for differences by race in the mean risk of rearrest. Column 4 adjusts for differences by race in the mean risk of rearrest for a violent crime. Robust standard errors, two-way clustered at the individual and judge level, are obtained by a bootstrapping procedure and appear in parentheses.

Appendix Table A12: Mean Risk and Unwarranted Disparity Estimates, Alternative Judge Decisions

	Linear Extrapolation	Quadratic Extrapolation	Local Linear Extrapolation
Panel A: Mean Risk by Race	(1)	(2)	(3)
White Defendants	0.343	0.341	0.345
	(0.007)	(0.027)	(0.033)
Black Defendants	0.405	0.415	0.447
	(0.006)	(0.022)	(0.038)
Panel B: System-Wide Discrimination			
Mean Across Cases	0.045	0.042	0.032
	(0.002)	(0.007)	(0.013)
Panel C: Judge-Level Discrimination			
Mean Across Judges	0.044	0.042	0.032
	(0.003)	(0.007)	(0.013)
Std. Dev. Across Judges	0.043	0.043	0.043
	(0.004)	(0.004)	(0.004)
Fraction Positive	0.855	0.838	0.769
	(0.018)	(0.043)	(0.081)
Judges	268	268	268

Notes. This table summarizes estimates of mean risk and unwarranted racial disparities from different extrapolations of the variation in Figure 2. The judge's decision variable in this table is release on recognizance (ROR) versus the assignment of any monetary bail, where there is a 5.8 percentage point disparity in the assignment of ROR between white and Black defendants after controlling for court-by-time effects. Panel A reports estimates of race-specific average misconduct risk, Panel B reports estimates of system-wide (case-weighted) unwarranted disparity, and Panel C reports empirical Bayes estimates of summary statistics for the judge-level unwarranted disparity prior distribution. To estimate mean risk, column 1 uses a linear extrapolation of the variation in Figure 2, while column 2 uses a quadratic extrapolation and column 3 uses a local linear extrapolation with a Gaussian kernel and a rule-of-thumb bandwidth. Robust standard errors, two-way clustered at the individual and judge level, are obtained by a bootstrapping procedure and appear in parentheses.

Appendix Table A13: Mean Risk and Unwarranted Disparity Estimates, Alternative Race Definition

	Linear Extrapolation	Quadratic Extrapolation	Local Linear Extrapolation
Panel A: Mean Risk by Race	(1)	(2)	(3)
White Defendants	0.208	0.138	0.187
	(0.009)	(0.018)	(0.014)
Black or Hispanic Defendants	0.393	0.419	0.415
	(0.005)	(0.019)	(0.011)
Panel B: System-Wide Discrimination			
Mean Across Cases	0.089	0.213	0.112
	(0.007)	(0.032)	(0.017)
Panel C: Judge-Level Discrimination			
Mean Across Judges	0.090	0.211	0.112
	(0.008)	(0.031)	(0.016)
Std. Dev. Across Judges	0.000	0.000	0.000
	(0.007)	(0.022)	(0.016)
Fraction Positive	1.000	1.000	1.000
	(0.018)	(0.004)	(0.015)
Judges	250	250	250

Notes. This table summarizes estimates of mean risk and unwarranted racial disparities from different extrapolations of the variation in Figure 2. The racial comparison in this table is between Black or Hispanic defendants to non-Hispanic white defendants, where there is a 8.4 percentage point release rate disparity after adjusting for court-by-time effects. Panel A reports estimates of race-specific average misconduct risk, Panel B reports estimates of system-wide (case-weighted) unwarranted disparity, and Panel C reports empirical Bayes estimates of summary statistics for the judge-level unwarranted disparity prior distribution. To estimate mean risk, column 1 uses a linear extrapolation of the variation in Figure 2, while column 2 uses a quadratic extrapolation and column 3 uses a local linear extrapolation with a Gaussian kernel and a rule-of-thumb bandwidth. Robust standard errors, two-way clustered at the individual and judge level, are obtained by a bootstrapping procedure and appear in parentheses.

Appendix Table A14: Mean Risk and Unwarranted Disparity Estimates by Defendant Characteristics

	Criminal History		Type of Arraignment Charge					
	Prior	No Prior	Felony	Misdemeanor	Drug	DUI	Property	Violent
Panel A: Mean Risk by Race	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
White Defendants	0.389	0.264	0.358	0.303	0.347	0.140	0.343	0.165
	(0.069)	(0.008)	(0.067)	(0.007)	(0.034)	(0.007)	(0.057)	(0.081)
Black Defendants	0.506	0.317	0.501	0.386	0.462	0.175	0.464	0.339
	(0.068)	(0.009)	(0.108)	(0.008)	(0.035)	(0.010)	(0.043)	(0.098)
Panel B: System-Wide Discrimination								
Mean Across Cases	0.026	0.024	0.030	0.046	0.055	0.024	0.010	0.107
	(0.024)	(0.002)	(0.058)	(0.003)	(0.009)	(0.005)	(0.020)	(1.678)
Panel C: Judge-Level Discrimination								
Mean Across Judges	0.026	0.024	0.030	0.046	0.058	0.024	0.006	0.105
	(0.024)	(0.002)	(0.057)	(0.003)	(0.009)	(0.005)	(0.020)	(1.648)
Std. Dev. Across Judges	0.038	0.014	0.034	0.035	0.052	0.000	0.036	0.031
	(0.010)	(0.004)	(0.027)	(0.003)	(0.007)	(0.010)	(0.014)	(1.072)
Fraction Positive	0.752	0.960	0.821	0.915	0.870	1.000	0.569	1.000
	(0.114)	(0.030)	(0.152)	(0.020)	(0.037)	(0.039)	(0.073)	(0.106)
Judges	263	264	261	264	258	174	222	219

Notes. This table summarizes estimates of mean risk and unwarranted racial disparities by defendant characteristics separately. For each subgroup, we require that a judge observe at least 25 cases in order to be included in the sample. Therefore, the number of judges does vary across the columns depending on how many judges in the sample meet the requirement. Information on demographics and criminal outcomes is derived from court records as described in the text. Prior is an indicator equal to one if the defendant has a prior conviction. Estimates come from a local linear extrapolation of the variation in Figure 2, although unlike Figure 2, the extrapolations are done within the given characteristic. Panel A reports estimates of race-specific average misconduct risk, Panel B reports estimates of system-wide (case-weighted) unwarranted disparity, and Panel C reports empirical Bayes estimates of summary statistics for the judge-level unwarranted disparity prior distribution. To estimate mean risk, this table uses a local linear extrapolation with a Gaussian kernel and a rule-of-thumb bandwidth. Robust standard errors, two-way clustered at the individual and judge level, are obtained by a bootstrapping procedure and appear in parentheses.

Appendix Table A15: Tests of Conventional MTE Monotonicity

	Number of Spline Knots			
	5	10	15	20
Panel A: White Defendants	(1)	(2)	(3)	(4)
Test Statistic	303.8	303.5	303.4	303.3
Deg. of Freedom	260	255	250	245
p-value	[0.032]	[0.020]	[0.012]	[0.007]
Cases	284,598	284,598	284,598	284,598
Panel B: Black Defendants				
Test Statistic	403.8	402.9	402.8	402.3
Deg. of Freedom	260	255	250	245
p-value	[<0.001]	[<0.001]	[<0.001]	[<0.001]
Cases	310,588	310,588	310,588	310,588

Notes. This table reports the results of the tests of conventional MTE monotonicity proposed by Frandsen et al. (2019), computed separately by defendant race. Test statistics are based on quadratic b-spline estimates of the relationship between misconduct outcomes and judge leniency, with the number of knots specified in each column, controlling for court-by-time fixed effects.

Appendix Table A16: Hierarchical MTE Model Hyperparameter Estimates

	White Defendants			Black Defendants		
	(1)	(2)	(3)	(4)	(5)	(6)
Mean Misconduct Risk (μ)	0.346	0.391	0.371	0.423	0.441	0.437
	(0.008)	(0.007)	(0.014)	(0.009)	(0.007)	(0.016)
Mean ln(Signal Quality) (α)	0.538	0.316	0.523	-0.038	-0.044	-0.080
	(0.128)	(0.074)	(0.125)	(0.146)	(0.075)	(0.104)
Mean Release Threshold (γ)	0.912	1.055	1.144	0.893	1.072	1.089
	(0.045)	(0.023)	(0.080)	(0.051)	(0.034)	(0.079)
Release Threshold Std. Dev. (δ)	0.369	0.109	0.149	0.417	0.194	0.203
	(0.039)	(0.011)	(0.037)	(0.052)	(0.021)	(0.049)
ln(Signal Quality) Std. Dev. (ψ)		0.140	0.134		0.166	0.151
		(0.019)	(0.016)		(0.014)	(0.013)
Regression of ln(Signal Quality)			-0.376			-0.007
on Release Threshold (β)			(0.153)			(0.212)
Judges	268	268	268	268	268	268

Notes. This table reports simulated minimum distance estimates of the MTE model described in the text. 500 simulation draws are used. Columns 3 and 6 estimate the full model with all hyperparameters. Columns 2 and 5 restrict $\beta = 0$, while columns 1 and 4 also restrict $\psi = 0$. The baseline model used in the text and summarized in Table 5 comes from columns 2 and 5 of this table. Robust standard errors, two-way clustered at the individual and the judge level, are reported in parentheses.

Appendix Table A17: Unwarranted Disparities and Judge Characteristics, Model-Based Mean Risk

	Full-Sample Disparities					Split-Sample Disparities	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
New Judge	-0.017				-0.013		-0.004
	(0.006)				(0.004)		(0.005)
Lenient Judge		-0.010			-0.012		-0.006
		(0.004)			(0.003)		(0.004)
Above-Median Black Share			-0.011		-0.007		0.000
			(0.004)		(0.005)		(0.006)
Manhattan Courtroom				0.036	0.033		0.024
				(0.005)	(0.004)		(0.005)
Bronx Courtroom				-0.004	-0.007		-0.002
				(0.004)	(0.005)		(0.007)
Queens Courtroom				0.028	0.022		0.020
				(0.005)	(0.006)		(0.007)
Richmond Courtroom				0.016	0.010		0.020
				(0.004)	(0.007)		(0.008)
Lagged Disparities						0.526	0.335
						(0.062)	(0.078)
Mean Disparity	0.049	0.049	0.049	0.049	0.049	0.049	0.049
R2	0.052	0.027	0.038	0.342	0.415	0.325	0.420
Judges	268	268	268	268	268	252	252

Notes. This table reports OLS estimates of regressions of unwarranted disparity posteriors on judge characteristics. Unwarranted disparities are estimated as described in Section 5, using the hierarchical MTE model estimate of mean risk. New judges are defined as judges appointed during our estimation period. Lenient judges are defined as judges with above-average leniency, controlling for court-by-time fixed effects. Courtroom locations are defined using the location of the modal case heard by each judge. Split-sample disparities are computed by splitting each judge's sample of cases at the median case and constructing two samples, a before-median case sample and an after-median case sample. Unwarranted disparities are then re-estimated within each subsample. The estimation procedure conditions on court-by-time effects, which causes a small number of judge effects to become collinear with the court-by-time effects and dropped. All specifications are weighted by the inverse variance of the unwarranted disparity posteriors. Robust standard errors are reported in parentheses.

Appendix Table A18: Marginal Treatment Effect Estimates of Racial Bias

	White Defendants	Black Defendants	Diff.
Panel A: Conventional MTE Estimates	(1)	(2)	(3)
Marginal Released Outcome	0.492	0.494	-0.003
	(0.027)	(0.025)	(0.030)
Panel B: IV Estimates			
Marginal Released Outcome	0.409	0.404	0.005
	(0.114)	(0.087)	(0.153)
Court x Time FE	Yes	Yes	–
Mean Misconduct	0.266	0.332	–
Cases	284,598	310,588	–

Notes. This table reports conventional MTE estimates and IV estimates of racial bias. The IV estimates instrument for pretrial release using a leave-one-out judge leniency measure. To estimate the MTE results, we first compute judge-specific release and misconduct rates that adjust for court-by-time fixed effects. We then fit a quadratic between misconduct rates and release rates. The MTE is equal to the derivative of this quadratic function. The regressions are estimated on the sample as described in the notes to Table 1. Standard errors are two-way clustered at the judge and defendant level. Standard errors are computed by a bootstrap procedure which resamples at the judge level with replacement and re-estimates the quadratic function between misconduct rates and release rates within each bootstrap sample. The standard error is equal to the standard deviation of the bootstrap estimates.

Appendix Table A19: Unwarranted Disparity Decompositions

	Baseline	No Racial Bias	Equal Signal Quality	Both
Panel A: Change Black Parameters	(1)	(2)	(3)	(4)
Unwarranted Disparity	0.047	-0.042	0.095	0.039
Release Rates (W/B)	0.768 / 0.703	0.768 / 0.795	0.768 / 0.652	0.768 / 0.709
Racial Bias	0.074	0.000	0.074	0.000
Marginal Outcomes (W/B)	0.650 / 0.577	0.650 / 0.650	0.650 / 0.577	0.650 / 0.650
Signal Quality (W/B)	1.386 / 0.970	1.386 / 0.970	1.386 / 1.386	1.386 / 1.386
Panel B: Change White Parameters				
Unwarranted Disparity		-0.006	0.136	0.062
Release Rates (W/B)		0.716 / 0.703	0.853 / 0.703	0.781 / 0.703
Racial Bias		0.000	0.074	0.000
Marginal Outcomes (W/B)		0.577 / 0.577	0.650 / 0.577	0.577 / 0.577
Signal Quality (W/B)		1.386 / 0.970	0.970 / 0.970	0.970 / 0.970
Judges	268	268	268	268

Notes. Column 1 of this table reports average unwarranted disparity and racial bias across judges and 250 simulations of the hierarchical MTE model, along with average release rates, marginal released outcomes, and signal quality of Black and white defendants. Simulations are based on the estimates from columns 2 and 4 of Appendix Table A16. Column 2 recomputes the statistics for a counterfactual in which Black (Panel A) or white (Panel B) release rates are set to eliminate racial bias, while column 3 adjusts Black (Panel A) or white (Panel B) signal quality to equalize signal quality across race. Column 4 applies both counterfactuals simultaneously.

Appendix Table A20: Model Estimates of Alternative Conditional Racial Disparities

	White Defendants	Black Defendants	Diff.
Disparity Conditional on:	(1)	(2)	(3)
Misconduct Potential	0.7591	0.7117	0.0474
	(0.0042)	(0.0065)	(0.0053)
Misconduct Signal	0.7360	0.7386	-0.0027
	(0.0064)	(0.0083)	(0.0112)
Misconduct Posterior	0.7811	0.7103	0.0707
	(0.0016)	(0.0019)	(0.0003)
Judges	268	268	–

Notes. This table reports estimates of average racial disparities in defendant release rates, conditional on different defendant unobservables. Estimates are given by the baseline hierarchical MTE model estimates and averages are taken both across judges and draws of the judge-level parameters. The first row conditions on true misconduct potential Y_i^*, yielding our unwarranted disparity measure. The second row conditions on the judge misconduct signal ν_{ij}, corresponding to the measure of race-blindness discussed in the text. The third row conditions on the judge misconduct posterior $p_j(\nu_{ij}, R_i)$ and captures racial bias. Robust standard errors, two-way clustered at the individual and the judge level, are reported in parentheses.

Appendix Table A21: Racial Bias and Judge Characteristics

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
New Judge	-0.021				-0.023		-0.007
	(0.009)				(0.007)		(0.005)
Lenient Judge		0.023			0.017		0.033
		(0.007)			(0.005)		(0.003)
Above-Median Black Share			-0.008		-0.013		-0.002
			(0.007)		(0.008)		(0.005)
Manhattan Courtroom				0.052	0.044		-0.006
				(0.008)	(0.008)		(0.006)
Bronx Courtroom				-0.016	-0.027		-0.015
				(0.007)	(0.010)		(0.006)
Queens Courtroom				0.038	0.023		-0.007
				(0.009)	(0.011)		(0.008)
Richmond Courtroom				0.037	0.019		-0.010
				(0.007)	(0.009)		(0.014)
Unwarranted Disparities						1.369	1.403
						(0.086)	(0.085)
Mean Bias	0.072	0.072	0.072	0.072	0.072	0.072	0.072
R2	0.026	0.053	0.007	0.332	0.397	0.646	0.770
Judges	268	268	268	268	268	268	268

Notes. This table reports OLS estimates of regressions of racial bias posteriors on judge characteristics. Posteriors are obtained from the heirarchical MTE model as described in Section 6. New judges are defined as judges appointed during our estimation period. Lenient judges are defined as judges with above-average leniency, controlling for court-by-time fixed effects. Courtroom locations are defined using the location of the modal case heard by each judge. All specifications are weighted by the inverse variance of the racial bias posteriors. Robust standard errors are reported in parentheses.

Appendix Table A22: Signal Quality Differences and Judge Characteristics

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
New Judge	-0.092				-0.073		-0.014
	(0.029)				(0.022)		(0.012)
Lenient Judge		0.031			0.016		0.074
		(0.020)			(0.016)		(0.009)
Above-Median Black Share			-0.031		-0.029		-0.006
			(0.020)		(0.024)		(0.013)
Manhattan Courtroom				0.172	0.153		-0.001
				(0.023)	(0.025)		(0.016)
Bronx Courtroom				-0.042	-0.062		-0.044
				(0.022)	(0.030)		(0.016)
Queens Courtroom				0.120	0.090		-0.018
				(0.028)	(0.034)		(0.021)
Richmond Courtroom				0.117	0.081		-0.050
				(0.023)	(0.029)		(0.037)
Unwarranted Disparities						4.575	4.584
						(0.197)	(0.215)
Mean Difference	0.412	0.412	0.412	0.412	0.412	0.412	0.412
R2	0.055	0.011	0.010	0.338	0.379	0.738	0.812
Judges	268	268	268	268	268	268	268

Notes. This table reports OLS estimates of regressions of differences in signal quality on judge characteristics. Posteriors are obtained from the heirarchical MTE model as described in Section 6. New judges are defined as judges appointed during our estimation period. Lenient judges are defined as judges with above-average leniency, controlling for court-by-time fixed effects. Courtroom locations are defined using the location of the modal case heard by each judge. All specifications are weighted by the inverse variance of the signal quality difference posteriors. Robust standard errors are reported in parentheses.

B Econometric Appendix

B.1 Defining and Measuring Discrimination with Multi-Valued Y_i^*

This appendix first generalizes our definition of racial discrimination and derivation of OVB in observational comparisons to settings where the decision-maker's objective is non-binary. We then discuss how our quasi-experimental framework for measuring racial discrimination extends to this case.

Our initial definition of racial discrimination, $\Delta_j = E[E[D_{ij} \mid Y_i^*, R_i = w] - E[D_{ij} \mid Y_i^*, R_i = b]]$, remains sensible in the case of non-binary Y_i^*, provided the support of Y_i^* is the same in the white ($R_i = w$) and Black ($R_i = b$) subpopulations. Natural generalizations of Equation (2) are given by

$$\Delta_j = \sum_{y \in Supp(Y_i^*)} \left(\delta_{jw}^y - \delta_{jb}^y\right) p_y \tag{B1}$$

in the multi-valued Y_i^* case, where $p_y = Pr(Y_i^* = y)$, and:

$$\Delta_j = \int_{Supp(Y_i^*)} \left(\delta_{jw}^y - \delta_{jb}^y\right) dF(y) \tag{B2}$$

in the case of continuous Y_i^*, where $F(\cdot)$ is the cumulative distribution function of Y_i^*. In both cases, $\delta_{jr}^y = E[D_{ij} \mid Y_i^* = y, R_i = r]$ gives conditional release rates for each race r and each $y \in Supp(Y_i^*)$.

As in Section 3.3, the bias of observational benchmarking regressions relative to these parameters, when judges are as-good-as-randomly assigned, is given by

$$\begin{aligned}\xi_j &= \sum_{y \in Supp(Y_i^*)} \delta_{jw}^y p_{yw} - \sum_{y \in Supp(Y_i^*)} \delta_{jb}^y p_{yb} - \sum_{y \in Supp(Y_i^*)} \left(\delta_{jw}^y - \delta_{jb}^y\right)(p_{yw} p_w + p_{yb} p_b) \\ &= \sum_{y \in Supp(Y_i^*)} \left(\delta_{jw}^y p_b + \delta_{jb}^y p_w\right)(p_{yw} - p_{yb})\end{aligned} \tag{B3}$$

in the multi-valued Y_i^* case, where $p_{yr} = Pr(Y_i^* = y \mid R_i = r)$ and again $p_r = Pr(R_i = r)$, and:

$$\begin{aligned}\xi_j &= \int_{Supp(Y_i^*)} \delta_{jw}^y dF_w(y) - \int_{Supp(Y_i^*)} \delta_{jb}^y dF_b(y) - \int_{Supp(Y_i^*)} \left(\delta_{jw}^y - \delta_{jb}^y\right) d(F_w(y) p_w + F_b(y) p_b) \\ &= \int_{Supp(Y_i^*)} \left(\delta_{jw}^y p_b + \delta_{jb}^y p_w\right) d(F_w(y) - F_b(y))\end{aligned} \tag{B4}$$

in the case of continuous Y_i^*, where $F_r(\cdot)$ is the cumulative distribution function of Y_i^* given $R_i = r$.

As in Section 5, discrimination is identified by the distribution of misconduct outcomes Y_i^* within each race when judges are quasi-randomly assigned. By Bayes' law:

$$\delta_{jr}^y = Pr(Y_i^* = y \mid D_{ij} = 1, R_i = r) \frac{E[D_{ij} \mid R_i = r]}{Pr(Y_i^* = y \mid R_i = r)} \tag{B5}$$

for multi-valued Y_i^* and similarly for continuous Y_i^*. The first two terms, $Pr(Y_i^* = y \mid D_{ij} = 1, R_i = r)$ and $E[D_{ij} \mid R_i = r]$, are identified by $Pr(Y_i = y \mid D_i = 1, Z_{ij} = 1, R_i = r)$ and $E[D_i \mid Z_{ij} = 1, R_i = r]$ under quasi-random judge assignment as before. In the continuous Y_i^* case, the first term is given by the conditional density of Y_i^* given $D_i = 1$, $Z_{ij} = 1$, and $R_i = r$. Estimates of the race-specific misconduct distribution corresponding to the third $Pr(Y_i^* = y \mid R_i = r)$ term (which might be

obtained from similar extrapolations of quasi-experimental data as in the binary Y_i^* case) thus yield a plug-in estimator of each δ_{jr}^y, which can be combined to estimate Δ_j according to the initial definitions.

B.2 Discrimination and Bias with Normally Distributed Signal Noise

This appendix derives the decision-making model discussed in Section 2. A judge observes noisy risk signals $\nu_i = Y_i^* + \eta_i$ with normally distributed noise: $\eta_i \mid Y_i^*, R_i \sim N(0, 1/\tau_{R_i}^2)$. The judge has potentially incorrect beliefs $\tilde{\mu}_r$ on race-specific average misconduct risk $\mu_r = E[Y_i^* \mid R_i = r]$ and knows the potentially race-specific quality of risk signals τ_r.

The judge's subjective posterior of misconduct risk, given a signal of $\nu_i = v$ for a defendant of race $R_i = r$, is derived from Bayes' rule:

$$\begin{aligned} p(\nu; r) &= \frac{\widetilde{Pr}(\nu_i = v \mid Y_i^* = 1, R_i = r)\widetilde{Pr}(Y_i^* = 1, R_i = r)}{\widetilde{Pr}(\nu_i = v, R_i = r)} \\ &= \frac{\phi(\tau_r(v-1))\tau_r\tilde{\mu}_r}{\phi(\tau_r(v-1))\tau_r\tilde{\mu}_r + \phi(\tau_r v)\tau_r(1-\tilde{\mu}_r)} \end{aligned} \tag{B6}$$

where $\widetilde{Pr}(\cdot)$ denotes subjective probabilities and $\phi(x) \propto \exp(-x^2/2)$ is the standard normal density. Simplifying, we have:

$$p(\nu; r) = \left(1 + \exp(\tau_r^2(1-2v)/2)\frac{1-\tilde{\mu}_r}{\tilde{\mu}_r}\right)^{-1} \tag{B7}$$

This specifies a risk-neutral judge's release rule, $D_i = \mathbf{1}[\pi_{R_i} \geq p(\nu_i; R_i)]$.

Equation (B7) shows that risk posteriors are strictly increasing in v, such that they can be inverted to write the judge's release decision as a cutoff rule for her observed signals ν_i:

$$D_i = \mathbf{1}\left[\frac{1}{2} - \ln\left(\frac{\tilde{\mu}_{R_i}(1-\pi_{R_i})}{\pi_{R_i}(1-\tilde{\mu}_{R_i})}\right)/\tau_{R_i}^2 \geq \nu_i\right] \tag{B8}$$

Equation (B8) shows that variation in risk beliefs $\tilde{\mu}_r$ and risk tolerances π_r are observationally equivalent in this model, in the sense that as one of these parameters varies in $(0,1)$ the other can be set to keep the index $I_r = \frac{\tilde{\mu}_r(1-\pi_r)}{\pi_r(1-\tilde{\mu}_r)}$, and thus the decision rule, constant.

A consequence of Equation (B8) is that the average misconduct rate of white and Black defendants at the margin of release, $E[Y_i^* \mid p(\nu_i; R_i) = \pi_r, R_i = r]$, is a function of the judges risk tolerance π_r and prior risk belief $\tilde{\mu}_r$. By Equation (4), the marginal outcomes under correct beliefs μ_r equals the judge's risk tolerance. More generally:

$$E[Y_i^* \mid p(\nu_i; R_i) = \pi_r, R_i = r] = \left(1 + I_r\left(\frac{1-\mu_r}{\mu_r}\right)\right)^{-1} \tag{B9}$$

by the observational equivalence of Equation (B8). Racial bias is found when this expression varies by race r, which could be due to racial animus ($\pi_w \neq \pi_b$) or inaccurate beliefs ($\tilde{\mu}_r \neq \mu_r$).

To characterize discrimination in this model, note that Equation (B8) and the conditional normal-

ity of ν_i implies that the judge's true and false negative rates can be written, respectively:

$$\delta_r^T = Pr(D_i = 1 \mid Y_i^* = 0, R_i = r) = \Phi\left(\frac{1}{2}\tau_r - \frac{1}{\tau_r}\ln I_r\right) \tag{B10}$$

$$\delta_r^F = Pr(D_i = 1 \mid Y_i^* = 1, R_i = r) = 1 - \Phi\left(\frac{1}{2}\tau_r + \frac{1}{\tau_r}\ln I_r\right) \tag{B11}$$

and the extent of racial discrimination is given by the extent to which $\Delta = (\delta_w^T - \delta_b^T)(1-\bar{\mu}) + (\delta_w^F - \delta_b^F)\bar{\mu}$ varies by race, for $\bar{\mu} = E[Y_i^*]$. With common signal quality, $\tau_w = \tau_b$, a lack of racial discrimination requires $I_w = I_b$. By comparison with Equation (B9), this scenario will generally lead to racial bias unless white and Black average misconduct risk are also equal ($\mu_w = \mu_b$). More generally, the fact that Δ is strictly decreasing (to zero) in the white index I_w and strictly increasing (to one) in the Black index I_b implies that there exist a set of thresholds (I_w, I_b) resulting in no racial discrimination on average, even when signal quality differs. Again, this will typically yield racial bias, per Equation (B9), to the extent either mean risk or signal quality differs by race.

B.3 Bail Release and Classification Error

This appendix shows how a judge minimizing the cost of type-I and type-II error in the bail setting implicitly uses a posterior risk threshold-crossing rule, as in Section 2. Suppose the cost of a type-I "false positive" decision (detaining an individual with no pretrial misconduct risk) is given by $c^I > 0$ and the cost of a type-II "false negative" decision (releasing an individual with pretrial misconduct risk) is given by $c^{II} > 0$. A judge's utility given a release decision $D_i \in \{0, 1\}$ is then:

$$U_i = -c^{II} D_i Y_i^* - c^I (1 - D_i)(1 - Y_i^*) \tag{B12}$$

Let $D(v)$ be a decision rule mapping risk signals ν_i to binary release decisions D_i. Suppose $D(v)$ is set to maximize the judge's expected utility (or minimize her expected disutility):

$$D(v) = \arg\min_{d(v)} \; c^{II} d(v) p(v) + c^I (1 - d(v))(1 - p(v)) \tag{B13}$$

where $p(\nu)$ denotes the judge's subjective expectation of pretrial misconduct given a signal of $\nu_i = \nu$. It is clear that this solution is a cutoff rule:

$$D(v) = \mathbf{1}[\pi \geq p(\nu_{ij})] \tag{B14}$$

where $\pi = \frac{c^{II}}{c^I + c^{II}} \in (0, 1)$ gives the judge's relative cost of type-II error. Per Equation (4), this also shows that when judge beliefs are accurate, the expected outcome of a marginally released defendant identifies this relative cost parameter.

B.4 Conventional Empirical Bayes Methods

This appendix summarizes the two conventional empirical Bayes approaches used in this paper: the posterior mean calculation of Morris (1983) and the posterior average effect calculation of Bonhomme and Weidner (2020). We use the former to plot the distribution of disparity posteriors in Figures 1, 3, and A3, and also to compute the prior means and standard deviations in these exhibits. We use the

latter to compute the fraction of judges with positive disparities in these figures, and also to interpret the coefficient estimates in Tables 4, A17, A21, and A22.

Let $\hat{\theta}_j$ be an estimate of an unknown judge-specific parameter θ_j, such as an observational benchmarking coefficient or our rescaled unwarranted disparity measure. Applying to the usual asymptotic approximation, we write $\hat{\theta}_j = \theta_j + \varepsilon_j$ where $\varepsilon_j \sim N(0, \Sigma_j)$ for known Σ_j. Conventional empirical Bayes methods further assume $\theta_j \sim N(\mu, \Lambda)$, where μ and Λ are unknown hyperparameters. Given this prior distribution, the posterior mean of θ_j after observing the estimate $\hat{\theta}_j$ is given by

$$E[\theta_j \mid \hat{\theta}_j] = \frac{\Sigma_j}{\Lambda + \Sigma_j}\mu + \frac{\Lambda}{\Lambda + \Sigma_j}\hat{\theta}_j \tag{B15}$$

More generally, Equation (B15) gives the minimum mean-squared error prediction of θ_j given $\hat{\theta}_j$ when the normality of θ_j is relaxed, provided μ and Λ continue to parameterize the mean and variance of the prior distribution.

Empirical Bayes posteriors estimate μ and Λ and plug these hyperparameter estimates into Equation (B15). We estimate μ and Λ by the weighted iterative procedure studied by (Morris, 1983), which is equivalent to a maximum likelihood procedure. At iteration k the hyperparameter estimates are:

$$\hat{\mu}_k = \sum_j \frac{\omega_{jk}}{\sum_{j'} \omega_{j'k}} \hat{\theta}_j \tag{B16}$$

$$\hat{\Lambda}_k = \sum_j \frac{\omega_{jk}}{\sum_{j'} \omega_{j'k}} \left((\hat{\theta}_j - \hat{\mu}_k)^2 - \Sigma_j \right) \tag{B17}$$

with inverse-variance weights that are proportional to $\omega_{jk} = (\hat{\Lambda}_{k-1} + \Sigma_j)^{-1}$ and where $\omega_{j0} = 1$. We iterate this procedure to convergence.

Bonhomme and Weidner (2020) discuss posterior average effect estimators of the cumulative distribution function for θ_j, given by

$$\hat{F}_\theta(t) = \frac{1}{J} \sum_j E[\mathbf{1}[\theta_j \leq t] \mid \hat{\theta}_j] \tag{B18}$$

for each t in the support of θ_j. Note that $1 - \hat{F}_\theta(0)$ is a posterior average effect estimate of the fraction of θ_j in the population that is positive. Under the normality assumption:

$$E[\mathbf{1}[\theta_j \leq t] \mid \hat{\theta}_j] = \Phi\left(-\frac{E[\hat{\theta}_j \mid \hat{\theta}_j]}{\sqrt{\frac{\Lambda \Sigma_j}{\Lambda + \Sigma_j}}} \right) \tag{B19}$$

which can, as with Equation (B15), be estimated by plugging in the estimates of the mean and variance hyperparameters. Just as with the empirical Bayes posterior estimator, Bonhomme and Weidner (2020) show that this posterior average effect estimator has certain robustness properties: it is optimal in terms of local worst-case bias, and its global bias is bounded by the minimum worst-case bias within a large class of estimators. They further show how regressions of the empirical Bayes posterior means on judge characteristics also have a posterior average effect interpretation and thus the same robustness properties for estimating conditional mean functions.

B.5 Bounding Mean Risk and Racial Discrimination

This appendix details the construction of mean risk and unwarranted disparity bounds in Appendix Table A9. As in the baseline analysis, this procedure uses estimates of race- and judge-specific release rates $\rho_{jr} = E[D_{ij} \mid R_i = r]$ and released misconduct rates $\lambda_{jr} = E[Y_i^* \mid D_{ij} = 1, R_i = r]$. Instead of extrapolating the latter estimates to estimate the mean risk parameters μ_{jr}, and the corresponding estimates of discrimination Δ_j, here we bound the range of logically possible μ_{jr} given typical misconduct rates of highly lenient judges and search within these ranges to bound statistics of the prior distribution of discrimination.

Each column of Appendix Table A9 forms bounds from a different leniency threshold $\overline{\rho} \in \{0.8, 0.85, 0.9\}$. For each race r, we first use a local linear regression of the estimated λ_{jr} on the estimated ρ_{jr} to estimate the average λ_{jr} for judges with $\rho_{jr} = \overline{\rho}$, parameters we denote by $\overline{\lambda}_r$. By definition, each $\overline{\lambda}_r$ bounds the mean risk of race r as

$$\mu_r \in [\overline{\lambda}_r \overline{\rho}, \overline{\lambda}_r \overline{\rho} + (1 - \overline{\rho})]. \tag{B20}$$

The lower bound $\overline{\lambda}_r \overline{\rho}$ is obtained from assuming all detained defendants for a judge with a leniency of $\overline{\rho}$ have $Y_i^* = 0$ while the upper bound is obtaind from assuming the $(1 - \overline{\rho})$ share of detained defendants have pretrial misconduct potential ($Y_i^* = 1$). Panel A of Appendix Table A9 reports estimates of these bounds for each race, along with their associated standard errors in parentheses. Note that by construction the width of each interval is equal to $1 - \overline{\rho}$.

To obtain bounds on the statistics in Panels B and C of Appendix Table A9, we perform grid searches within the mean risk bounds in Panel A. For example, to bound the system-wide level of discrimination we search within the mean risk bounds to find the (μ_w, μ_b) pair that minimizes and maximizes the case-weighted average of judge-specific unwarranted disparity Δ_j. We report these bounds and their associated standard errors in parentheses. Note that the width of each statistic's interval is weakly increasing in $1 - \overline{\rho}$, reflecting the increase in the range of mean risk parameters.

B.6 Conventional Monotonicity Violations and Judge Signal Quality

This appendix shows how differences in the way judges consider defendant and case characteristics, which lead to violations of conventional MTE monotonicity, can be viewed as differences in judge signal quality within models like the one we develop in Section 3.2. In doing so we show that such models are without observational loss, provided judge release decisions are better-than-random.

Consider a setting with a binary potential misconduct outcome Y_i^* and a set of binary judge release decisions D_{ij}. The distribution of these random variables is fully specified by the mean risk $\mu = E[Y_i^*]$ and the true and false negative rates $\delta_j^T = E[D_{ij} \mid Y_i^* = 0]$ and $\delta_j^F = E[D_{ij} \mid Y_i^* = 0]$. With mean risk fixed, any restriction on judicial decision-making – such as conventional MTE monotonicity or alternative parameterizations – can thus be understood as restricting the set of (δ_j^T, δ_j^F).

We first show that when judges are making better-than-random release decisions, in the sense of $0 < \delta_j^T < \delta_j^F < 1$ for each j, it is without observational loss to assume a decision-making model of $D_{ij} = \mathbf{1}[\kappa_j \geq Y_i^* + \eta_i/\tau_j]$, with $\eta_i \mid Y_i^*$ following a known continuous distribution and $\tau_j > 0$. This follows since then $\tau_j = G_\eta^{-1}(\delta_j^T) - G_\eta^{-1}(\delta_{1j}^F) > 0$ and $\kappa_j = G_\eta^{-1}(\delta_j^T)/\tau_j$ rationalize each (δ_j^T, δ_j^F), where

$G_\eta(\cdot)$ specifies the cumulative distribution of $\eta_i \mid Y_i^*$:

$$\begin{aligned} E[D_{ij} \mid Y_i^* = y] &= Pr(\kappa_j \geq y + \eta_i/\tau_j) \\ &= G_\eta((\kappa_j - y)\tau_j) \\ &= G_\eta(G_\eta^{-1}(\delta_j^T)) + y(G_\eta^{-1}(\delta_j^F) - G_\eta^{-1}(\delta_j^T)) \\ &= \delta_j^T + y(\delta_j^F - \delta_j^T) \end{aligned} \tag{B21}$$

In particular, Equation (B21) shows that our risk signal threshold decision rule (23), in which $\eta_i \mid Y_i^* \sim N(0,1)$, is without loss in this case. In general, we may think of τ_j as capturing judge j's signal quality: how less likely she is to release defendants with $Y_i^* = 1$ than those with $Y_i^* = 0$.

We next relate differences in such signal quality to conventional monotonicity violations in a simple behavioral model of judicial decision-making. Suppose judges observe a vector of defendant and case characteristics X_i^* which are, without loss, mean zero and positively correlated with misconduct potential: $\mu_X(1) \equiv E[X_i^* \mid Y_i^* = 1] > E[X_i^* \mid Y_i^* = 0] \equiv \mu_X(0)$. Judges place different weights β_j on the elements of this vector and also vary in their overall leniency π_j, such that:

$$D_{ij} = \mathbf{1}[\pi_j \geq X_i^{*\prime}\beta_j + U_i] \tag{B22}$$

where we assume $U_i \mid X_i^*, Y_i^*$ is uniformly distributed. In this model $E[D_{ij} \mid Y_i^* = y] = \pi_j - \mu_X(y)'\beta_j$, assuming the parameters are such that these are all between zero and one.

Conventional monotonicity in this model requires $Pr(D_{ij} \geq D_{ik} = 1)$ or $Pr(D_{ik} \geq D_{ij} = 1)$ for each (j,k), which generally restricts the weights β_j to be the same across judges. If some elements of X_i^* were observed to the econometrician, one could relax this assumption by a conditional analysis within sets of defendants with identical observables (e.g., Mueller-Smith, 2015). Conditional monotonicity would then generally constrain the weights corresponding to unobserved characteristics to be constant.

Judicial decision-making is here better-than-random when $\delta_j^T - \delta_j^F = (\mu_X(1) - \mu_X(0))'\beta_j > 0$ or when the weights in each β_j are non-negative with at least one element strictly positive. In this case we have from the above result an equivalent representation of:

$$D_{ij} = \mathbf{1}[\kappa_j \geq Y_i^* + V_i/\tau_j] \tag{B23}$$

where $V_i \mid Y_i^* \sim U(0,1)$. Here judge signal quality is given by $\tau_j = (\mu_X(1) - \mu_X(0))'\beta_j$ and has an straightforward interpretation: with only one element in X_i^*, for example, differences in τ_j are proportional to differences in the behavioral weights β_j. More generally, this discussion shows how parameterizations of the distribution of signal quality across judges can be thought to structure differences in how judges weigh defendant and case characteristics when making release decisions.

B.7 SMD Estimation of the Hierarchical MTE Model

We estimate the hierarchical model described in Section 6.1 and Appendix B.2 by a simulated minimum distance (SMD) procedure that targets moments of the distribution of race- and judge-specific release rates $\rho_{jr} = E[D_{ij} \mid R_i = r]$ and released misconduct rates $\lambda_{jr} = E[Y_i^* \mid D_{ij} = 1, R_i = r]$, estimated from quasi-experimental judge assignments. This appendix formally specifies this procedure.

We first obtain estimates of ρ_{jr} and λ_{jr} from OLS regressions of pretrial release D_i and pretrial

misconduct Y_i on judge-by-race interactions, adjusting for the quasi-experimental court-by-time effects) and defendant and case observables as discussed in Section 5.2. Subject to the usual asymptotic approximation, the resulting estimates $\hat{\rho}_{jr}$ and $\hat{\lambda}_{jr}$ can be modeled as noisy measures of the true parameters, with a known distribution of sampling error. Specifically:

$$\hat{\rho}_{jr} = \rho_{jr} + \varepsilon^{\rho}_{jr} \tag{B24}$$

$$\hat{\lambda}_{jr} = \lambda_{jr} + \varepsilon^{\lambda}_{jr} \tag{B25}$$

where $\varepsilon \mid \rho, \lambda \sim N(0, \Sigma)$ for a variance-covariance matrix Σ that is given by conventional asymptotics. Let $\mathcal{X} = ((\hat{\rho}_{jr}, \hat{\lambda}_{jr})_{j=1,\ldots,268, r\in\{w,b\}})$ collect these estimates across the 268 judges in our sample and both races w and b.

The model in Appendix B.2 specifies ρ_{jr} and λ_{jr} as functions of mean misconduct risk μ_r, judge signal quality τ_{jr}, and risk thresholds π_{jr}:

$$\rho_{jr} = \Phi((f(\pi_{jr}, \mu_r, \tau_{jr}) - 1)\tau_{jr}))\mu_r + \Phi(f(\pi_{jr}, \mu_r, \tau_{jr})\tau_{jr}))(1 - \mu_r) \tag{B26}$$

$$\lambda_{jr} = \Phi((f(\pi_{jr}, \mu_r, \tau_{jr}) - 1)\tau_{jr}))\mu_r / \rho_{jr} \tag{B27}$$

where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function and $f(\cdot)$ is as defined in Section B.2. We further model signal thresholds $\kappa_{jr} = f(\pi_{jr}, \mu_r, \tau_{jr})$ and log signal quality $\ln \tau_{jr}$ as being joint-normally distributed across judges, with reisdual correlation across races. That is, we specify:

$$\ln \tau_{jr} = \alpha_r + \beta_r \kappa_{jr} + \epsilon_{jr} \tag{B28}$$

for each race r, with $(\kappa_{jw}, \kappa_{jb})' \sim N(\mu_\kappa, \Lambda_\kappa)$ and $(\epsilon_{jw}, \epsilon_{jb})' \mid \kappa \sim N(0, \Lambda_\tau)$.

Equations (B24)–(B28) specify a complete distribution for the observed quasi-experimental estimates $\mathcal{X}$ in terms of a hyperparameter vector $\Theta = (\mu_w, \mu_b, \alpha_w, \alpha_b, \beta_w, \beta_b, \mu'_\kappa, vec(\Lambda_\kappa^{1/2})', vec(\Lambda_\tau^{1/2})')'$. In practice, there is no simple closed form expression for this likelihood, complicating maximum likelihood estimation. Instead, we estimate Θ by SMD, targeting moments of $\mathcal{X}$ as motivated by the discussion in Section 6.1. Specifically, let $\hat{M}$ be a vector with the first two race-specific elements of:

$$\hat{M}_{1r} = \sum_{j=1}^{268} \omega^{\rho}_{jr} \hat{\rho}_{jr} \tag{B29}$$

$$\hat{M}_{2r} = \sum_{j=1}^{268} \omega^{\rho}_{jr} (\hat{\rho}_{jr} - \hat{M}_{1r})^2 \tag{B30}$$

the next three race-specific elements corresponding to coefficient estimates from the ω^{λ}_{jr}-weighted quadratic OLS regression of:

$$\hat{\lambda}_{jr} = \hat{M}_{3r} + \hat{M}_{4r}\hat{\rho}_{jr} + \hat{M}_{5r}\hat{\rho}^2_{jr} + \hat{\upsilon}_{jr} \tag{B31}$$

and the sixth race-specific element corresponding to the ω^{λ}_{jr}-weighted residual variance estimate:

$$\hat{M}_{6r} = \sum_{j=1}^{268} \omega^{\lambda}_{jr} \hat{\upsilon}^2_{jr} \tag{B32}$$

The weights are derived from the estimation error matrix Σ: ω^{ρ}_{jr} is proportional to the inverse variance of $\hat{\rho}_{jr} - \rho_{jr}$ while ω^{λ}_{jr} is proportional to the inverse variance of $\hat{\lambda}_{jr} - \lambda_{jr}$, with both weights rescaled to sum to one in the population of judges. We further include in $\hat{M}$ the $\sqrt{\omega^{\rho}_{jw}\omega^{\rho}_{jb}}$-weighted covariance of $\hat{\rho}_{jw}$ and $\hat{\rho}_{jw}$ as well as the $\sqrt{\omega^{\lambda}_{jw}\omega^{\lambda}_{jb}}$-weighted covariance of $\hat{\lambda}_{jw}$ and $\hat{\lambda}_{jw}$. Together this gives 14 elements in $\hat{M}$, the same number of hyperparameters in Θ.

To estimate Θ we use an SMD procedure that matches the empirical moments in $\hat{M}$ with the corresponding model-implied moments averaged across 500 simulated draws of the above data-generating process. That is, we estimate:

$$\hat{\Theta} = \arg\min_{\Theta} \sum_{m=1}^{14} \left(\hat{M}_m - \frac{1}{500} \sum_{s=1}^{500} M_{ms}(\Theta) \right)^2 \tag{B33}$$

where the functions $M_{ms}(\cdot)$ of candidate hyperparameters Θ are given by applying the previous moment calculations to data generated from 500 fixed simulation draws s. Conventional asymptotic theory for $\hat{\Theta}$ applies under appropriate regularity conditions (e.g., Pakes and Pollard, 1989).

Columns 3 and 6 of Appendix Table A16 report SMD estimates and standard errors for the full model. As discussed in the main text, our baseline model estimates set $\beta_r = 0$. Per the intuition in Section 6.1 and to keep the model just-identified, we correspondingly drop the quadratic term from the moment regression in Equation (B31). The resulting estimates are reported in columns 2 and 5 of Appendix Table A16. To impose conventional MTE monotonicity, we further set the variance of τ_{jr} to zero. The resulting estimates are reported in columns 1 and 4 of Appendix Table A16.

Lastly, given $\hat{\Theta}$, we compute maximum *a posteriori* probability estimates (also known as posterior modes) of the judge-specific parameters $\theta_j = (\kappa_{jw}, \ln \tau_{jw}, \kappa_{jb}, \ln \tau_{jb})'$, following an approach similar to that which Angrist et al. (2017) apply for a similar hierarchical model. Note that the log-likelihood of $\theta = (\theta_1' \ldots, \theta_{268}')'$ and quasi-experimental estimates $\mathcal{X}$ can be written:

$$\mathcal{L}(\theta, \mathcal{X}) = \ln \phi_m \left(\mathcal{X} - \bar{X}(\theta); \Sigma \right) + \ln \phi_m \left(\theta - \mu_\theta; \Lambda_\theta \right) \tag{B34}$$

where $\phi_m(\cdot; V)$ gives the density of a mean-zero multivariate normal vector with variance-covariance matrix V; $\bar{X}(\cdot)$ collects the formulas from Equations (B26) and (B27), for ρ_{jr} and λ_{jr} in terms of μ_w, μ_b, and θ; and both μ_θ and Λ_θ are derived from the α_r and β_r, μ_κ, Λ_κ, and Λ_τ. Our estimates of θ are given by maximizing this likelihood, plugging in our baseline hyperparameter estimates $\hat{\Theta}$.