ONLINE APPENDIX – RACIAL BIAS IN BAIL DECISIONS

David Arnold* Will Dobbie† Crystal S. Yang‡

September 2020

Contents

*Princeton University. Email: dharnold@princeton.edu
†Princeton University and NBER. Email: wdobbie@princeton.edu
‡Harvard Law School and NBER. Email: cyang@law.harvard.edu

Corrections to Published Paper

This appendix makes precise the formal definition of racial bias in our article "Racial Bias in Bail Decisions" published in the *Quarterly Journal of Economics* in November 2018. Our paper defines a judge as racially biased if her decisions cannot be solely explained by accurate statistical discrimination. Therefore, a judge is racially biased if she perceives a higher threshold of release for black defendants than white defendants at the margin, or under an alternative model, if she overestimates the cost of release for black defendants relative to white defendants at the margin. We refer to this verbal definition repeatedly throughout the paper. However, our formal definition of bias was insufficiently precise as to our intended definition of bias, which we realized in light of a working paper by Canay, Mogstad, and Mountjoy (2020).

To make our intended definition of bias clear, we make the following amendments to the published paper, where page numbers refer to the published version:

(1) On p. 1893, at the end of the paragraph beginning "The perceived benefit of release for defendant i..." we add the following definitions: "Let the non-race characteristics of the marginal defendant for judge j and race r be denoted $\mathbf{V}^*_{i,r}$. We correspondingly define $t^{j*}_r = t^j_r(\mathbf{V}^*_{i,r})$."

(2) On p. 1893, Definition 1 should be: "DEFINITION 1. Following Becker (1957, 1993), we define judge j as racially biased against black defendants if $t^{j*}_W > t^{j*}_B$. Thus, for racially biased judges, there is a higher perceived benefit of releasing white defendants than black defendants at the margin."

(3) On p. 1894, at the end of the sentence beginning "Given this decision rule...," $t^j_r(\mathbf{V}_i)$ should be $t^j_r(\mathbf{V}^*_{i,r})$ and at the end of the sentence beginning "We simplify our notation...," α^j_r should be $\alpha^j_r = \mathbb{E}[\alpha^j_i | \mathbf{V}_i = \mathbf{V}^*_{i,r}, r_i = r]$.

(4) On p. 1895, Definition 2 should be: "DEFINITION 2. We define judge j as making racially biased prediction errors in risk against black defendants if $\tau^j_W(\mathbf{V}_i = \mathbf{V}^*_{i,W}) > \tau^j_B(\mathbf{V}_i = \mathbf{V}^*_{i,B})$. Thus, judges making racially biased prediction errors systematically overestimate the true cost of release for black defendants relative to white defendants at the margin."

(5) In Equations (4), (5), (6), (8) and any discussion of these equations, t^j_r should be t^{j*}_r.

(6) On p. 1922, in the sentence beginning with "Bail judges could, for example, harbor..." the phrase "observably similar" should be struck.

(7) We have made the corresponding notational changes in the online appendix that follows.

Online Appendix A: Additional Results

Online Appendix Table A1
Racial Bias in the Assignment of Non-Monetary Bail

	White	Black	D^{IV}
Panel A: Pre-Trial Release	(1)	(2)	(3)
Pre-trial Release	0.490***	0.511***	−0.021
	(0.081)	(0.045)	(0.092)
	[0.711]	[0.688]	–
Panel B: Pre-Trial Misconduct			
Rearrest Prior to Disposition	0.085*	−0.009	0.094
	(0.050)	(0.039)	(0.065)
	[0.172]	[0.182]	–
Rearrest for Drug Crime	0.060**	−0.026	0.086**
	(0.030)	(0.026)	(0.041)
	[0.077]	[0.081]	–
Rearrest for Property Crime	0.087**	0.001	0.086*
	(0.037)	(0.029)	(0.048)
	[0.065]	[0.068]	–
Rearrest for Violent Crime	0.033	0.010	0.022
	(0.029)	(0.027)	(0.040)
	[0.047]	[0.071]	–
Observations	106,846	149,407	–

Notes. This table reports estimates of the impact of assigning non-monetary bail (defined as both ROR and non-monetary conditions) versus monetary bail on pre-trial release (Panel A) and pre-trial misconduct (Panel B). Columns (1)–(2) report two-stage least squares results of the impact of pre-trial release on the probability of pre-trial misconduct separately by race, while column (3) reports the difference between the white and black two-stage least squares coefficients, or D^{IV} as described in the text. All specifications use IV weights for each specification and report robust standard errors two-way clustered at the individual and judge-by-shift level in parentheses. All specifications also control for court-by-time fixed effects and defendant race, gender, age, whether the defendant had a prior offense in the past year, whether the defendant had a prior history of pre-trial crime in the past year, whether the defendant had a prior history of failure to appear in the past year, the number of charged offenses, indicators for crime type (drug, DUI, property, violent, or other), crime severity (felony or misdemeanor), and indicators for any missing characteristics. The sample means of the dependent variables are reported in brackets. *** = significant at 1 percent level, ** = significant at 5 percent level, * = significant at 10 percent level.

Online Appendix Table A2
White-Hispanic Bias in Pre-Trial Release

	IV Results			MTE Results		
	White	Hispanic	D^{IV}	White	Hispanic	D^{MTE}
Panel A: Rearrest for All Crimes	(1)	(2)	(3)	(4)	(5)	(6)
Rearrest Prior to Disposition	0.274***	0.246**	0.028	0.299**	0.260**	0.039
	(0.099)	(0.119)	(0.147)	(0.129)	(0.129)	(0.184)
	[0.167]	[0.176]	–	[0.167]	[0.176]	–
Panel B: Rearrest by Crime Type						
Rearrest for Drug Crime	0.098	0.079	0.020	0.084	0.098	−0.014
	(0.067)	(0.068)	(0.093)	(0.086)	(0.075)	(0.115)
	[0.066]	[0.087]	–	[0.066]	[0.087]	–
Rearrest for Property Crime	0.117	0.211**	−0.094	0.140	0.166	−0.026
	(0.073)	(0.102)	(0.125)	(0.102)	(0.115)	(0.156)
	[0.066]	[0.064]	–	[0.066]	[0.064]	–
Rearrest for Violent Crime	0.012	0.141**	−0.129	0.008	0.147**	−0.139
	(0.052)	(0.069)	(0.088)	(0.070)	(0.072)	(0.102)
	[0.043]	[0.052]	–	[0.043]	[0.052]	–
Observations	35,914	48,447	–	35,914	48,447	–

Notes. This table reports estimates of white non-Hispanic versus white-Hispanic bias in pre-trial release based on rearrest prior to case disposition. Columns (1)–(2) report two-stage least squares results of the impact of pre-trial release on the probability of pre-trial misconduct separately by race, while column (3) reports the difference between the white non-Hispanic and white Hispanic two-stage least squares coefficients, or D^{IV} as described in the text. Columns (1)–(3) use IV weights for each specification and report robust standard errors two-way clustered at the individual and judge-by-shift level in parentheses. Columns (4)–(5) report the average marginal treatment effect of the impact of pre-trial release on the probability of pre-trial misconduct separately by race, while column (6) reports the difference between the white non-Hispanic and white Hispanic MTE coefficients, or D^{MTE} as described in the text. Columns (4)–(6) use equal weights for each judge and report bootstrapped standard errors clustered at the judge-by-shift level in parentheses. All specifications control for court-by-time fixed effects and defendant race, gender, age, whether the defendant had a prior offense in the past year, whether the defendant had a prior history of pre-trial crime in the past year, whether the defendant had a prior history of failure to appear in the past year, the number of charged offenses, indicators for crime type (drug, DUI, property, violent, or other), crime severity (felony or misdemeanor), and indicators for any missing characteristics. The sample means of the dependent variables are reported in brackets. *** = significant at 1 percent level, ** = significant at 5 percent level, * = significant at 10 percent level.

Online Appendix Table A3

First Stage Results by Case Characteristics

	Crime Severity		Crime Type			Defendant Type	
	Misd.	Felony	Property	Drug	Violent	Prior	No Prior
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Pre-trial Release	0.584***	0.204***	0.516***	0.364***	0.119***	0.452***	0.346***
	(0.042)	(0.035)	(0.046)	(0.048)	(0.041)	(0.038)	(0.028)
	[0.721]	[0.674]	[0.607]	[0.785]	[0.685]	[0.587]	[0.587]
Court x Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Crime Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	128,409	127,844	55,432	83,277	74,193	87,424	168,829

Notes. This table reports the first stage relationship between pre-trial release and judge leniency in different subsamples. The regressions are estimated on the sample as described in the notes to Table 1. Judge leniency is estimated using data from other cases assigned to a bail judge in the same year, constructed separately by defendant race, following the procedure described in Section II.B. All regressions include court-by-time fixed effects and baseline controls for race, gender, age, whether the defendant had a prior offense in the past year, whether the defendant had a prior history of pre-trial crime in the past year, whether the defendant had a prior history of failure to appear in the past year, the number of charged offenses, indicators for crime type (drug, DUI, property, violent, and other), crime severity (felony and misdemeanor), and indicators for any missing controls. The sample mean of the dependent variable is reported in brackets. Robust standard errors two-way clustered at the individual and judge-by-shift level are reported in parentheses. *** = significant at 1 percent level, ** = significant at 5 percent level, * = significant at 10 percent level.

Online Appendix Table A4
OLS Results

	White	Black	Difference
Panel A: Rearrest for All Crimes	(1)	(2)	(3)
Rearrest Prior to Disposition	0.181***	0.188***	−0.007**
	(0.003)	(0.002)	(0.004)
	[0.172]	[0.182]	–
Panel B: Rearrest by Crime Type			
Rearrest for Drug Crime	0.097***	0.103***	−0.006**
	(0.002)	(0.002)	(0.002)
	[0.077]	[0.081]	–
Rearrest for Property Crime	0.067***	0.073***	−0.006*
	(0.002)	(0.002)	(0.003)
	[0.065]	[0.068]	–
Rearrest for Violent Crime	0.052***	0.063***	−0.010***
	(0.002)	(0.002)	(0.002)
	[0.047]	[0.071]	–
Observations	106,846	149,407	–

Notes. This table reports OLS results of racial bias in pre-trial release based on rearrest prior to case disposition. The regressions are estimated on the sample as described in the notes to Table 1. Columns (1)–(2) report OLS estimates of the impact of pre-trial release on the probability of pre-trial misconduct separately by race, while column (3) reports the difference between the white and black OLS coefficients. Robust standard errors two-way clustered at the individual and judge-by-shift level are reported in parentheses. The sample means of the dependent variables are reported in brackets. All specifications control for court-by-time fixed effects and defendant race, gender, age, whether the defendant had a prior offense in the past year, whether the defendant had a prior history of pre-trial crime in the past year, whether the defendant had a prior history of failure to appear in the past year, the number of charged offenses, indicators for crime type (drug, DUI, property, violent, or other), crime severity (felony or misdemeanor), and indicators for any missing characteristics. *** = significant at 1 percent level, ** = significant at 5 percent level, * = significant at 10 percent level.

Online Appendix Table A5

Results for Other Definitions of Pre-Trial Misconduct

	Philadelphia		Miami		Pooled	
	D^{IV}	D^{MTE}	D^{IV}	D^{MTE}	D^{IV}	D^{MTE}
	(1)	(2)	(3)	(4)	(5)	(6)
Rearrest	0.045	0.078	0.263**	0.249**	0.222**	0.231**
	(0.183)	(0.194)	(0.115)	(0.121)	(0.101)	(0.117)
	[0.194]	[0.194]	[0.149]	[0.149]	[0.178]	[0.178]
FTA	−0.024	0.006	–	–	–	–
	(0.187)	(0.202)				
	[0.204]	[0.204]	–	–	–	–
FTA or Rearrest	0.008	0.042	0.263**	0.249**	0.208**	0.314*
	(0.209)	(0.221)	(0.115)	(0.121)	(0.102)	(0.189)
	[0.318]	[0.318]	[0.149]	[0.149]	[0.256]	[0.256]
Observations	162,836	162,836	93,417	93,417	256,253	256,253

Notes. This table reports estimates of racial bias in pre-trial release based on rearrest prior to case disposition, FTA (available only in Philadelphia), and either rearrest or FTA. Columns (1)–(2) report two-stage least squares estimates of D^{IV} and MTE estimates of D^{MTE} for Philadelphia. Columns (3)–(4) report two-stage least squares estimates of D^{IV} and MTE estimates of D^{MTE} for Miami. Columns (5)–(6) report two-stage least squares estimates of D^{IV} and MTE estimates of D^{MTE} for the pooled sample. For IV specifications, robust standard errors two-way clustered at the individual and judge-by-shift level reported in parentheses. For MTE specifications, bootstrapped standard errors clustered at the judge-by-shift level are reported in parentheses. All specifications control for court-by-time fixed effects and defendant race, gender, age, whether the defendant had a prior offense in the past year, whether the defendant had a prior history of pre-trial crime in the past year, whether the defendant had a prior history of failure to appear in the past year, the number of charged offenses, indicators for crime type (drug, DUI, property, violent, or other), crime severity (felony or misdemeanor), and indicators for any missing characteristics. The sample means of the dependent variables are reported in brackets. *** = significant at 1 percent level, ** = significant at 5 percent level, * = significant at 10 percent level.

Online Appendix Table A6
Social Cost of Crime Results

	D^{IV} Estimate	D^{MTE} Estimate	Lower Bound	Upper Bound
	(1)	(2)	(3)	(4)
Rearrest for Robbery	0.028	0.035	$73,196	$333,701
	(0.034)	(0.037)		
Rearrest for Assault	0.068	0.065	$41,046	$109,903
	(0.050)	(0.057)		
Rearrest for Burglary	0.047	0.018	$50,291	$50,291
	(0.048)	(0.058)		
Rearrest for Theft	0.118*	0.081	$9,598	$9,974
	(0.062)	(0.075)		
Rearrest for Drug	0.047	0.097	$2,544	$2,544
	(0.060)	(0.067)		
Rearrest for DUI	0.007	0.016	$25,842	$25,842
	(0.009)	(0.012)		

Notes. This table reports the difference in two-stage least squares and marginal treatment effect estimates of the impact of pre-trial release on the probability of pre-trial misconduct between white and black defendants for different crimes. We exclude rearrest for crime types that are extremely rare, e.g. murder and rape, and crime types that cannot be categorized into the listed categories, e.g. disorderly conduct. The regressions are estimated on the sample as described in the notes to Table 1. The dependent variable is listed in each row. In column (1), robust standard errors two-way clustered at the individual and judge-by-shift level are reported in parentheses. In column (2), bootstrap standard errors clustered at the judge-by-shift level are reported in parentheses. All specifications control for court-by-time fixed effects and defendant race, gender, age, whether the defendant had a prior offense in the past year, whether the defendant had a prior history of pre-trial crime in the past year, whether the defendant had a prior history of failure to appear in the past year, the number of charged offenses, indicators for crime type (drug, DUI, property, violent, or other), crime severity (felony or misdemeanor), and indicators for any missing characteristics. *** = significant at 1 percent level, ** = significant at 5 percent level, * = significant at 10 percent level.

Online Appendix Table A7
Robustness Results

	Estimates of D^{IV}					Estimates of D^{MTE}		
	Drop Impossible	Re-Weight Char.	Drop Hispanics	Cluster by Judge	Control Bail $	Drop Impossible	Drop Hispanics	Cluster by Judge
Panel A: Rearrest for All Crimes	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Rearrest Prior to Disposition	0.215**	0.238**	0.238**	0.222**	0.252**	0.221**	0.259*	0.231*
	(0.095)	(0.103)	(0.119)	(0.102)	(0.104)	(0.098)	(0.151)	(0.126)
Panel B: Rearrest by Crime Type								
Drug Crime	0.059	0.055	0.066	0.047	0.054	0.109	0.094	0.097
	(0.059)	(0.057)	(0.080)	(0.061)	(0.062)	(0.069)	(0.097)	(0.073)
Property Crime	0.150**	0.181**	0.105	0.163**	0.178**	0.096	0.074	0.106
	(0.066)	(0.076)	(0.083)	(0.077)	(0.076)	(0.074)	(0.106)	(0.097)
Violent Crime	0.059	0.103*	0.006	0.080	0.100	0.048	−0.004	0.083
	(0.054)	(0.060)	(0.069)	(0.064)	(0.062)	(0.062)	(0.088)	(0.070)
Observations	252,992	256,253	170,923	256,253	256,253	252,992	170,923	256,253

Notes. This table reports robustness checks for our estimates of D^{IV} and D^{MTE}. Columns (1) and (6) drop the four percent of cases where defendants are reported as being detained but are rearrested prior to disposition. Column (2) re-weights cases so that the white and black samples have identical observable characteristics following the procedure described in Appendix B. Columns (3) and (7) drop Hispanic whites from the sample. Column (4) reports standard errors clustered by defendant and judge. Column (5) instruments for monetary bail amount with a leave-out leniency measure constructed using monetary bail amount. Column (8) reports standard errors clustered by judge. All regressions include court-by-time fixed effects and baseline controls for race, gender, age, whether the defendant had a prior offense in the past year, whether the defendant had a prior history of pre-trial crime in the past year, whether the defendant had a prior history of failure to appear in the past year, the number of charged offenses, indicators for crime type (drug, DUI, property, violent, and other), crime severity (felony and misdemeanor), and indicators for any missing controls. *** = significant at 1 percent level, ** = significant at 5 percent level, * = significant at 10 percent level.

Online Appendix Table A8

Mean Pre-Trial Release and Misconduct Rates by Judge and Defendant Race

	Race of Judge	
	White	Black
Panel A: Pre-Trial Release Rates	(1)	(2)
White Defendant Release Rate	0.557	0.552
	(0.497)	(0.497)
Black Defendant Release Rate	0.535	0.530
	(0.499)	(0.499)
Panel B: Pre-Trial Rearrest Rates		
White Defendant Rearrest Rate	0.207	0.202
	(0.405)	(0.402)
Black Defendant Rearrest Rate	0.280	0.294
	(0.449)	(0.456)

Notes. This table presents mean rates of pre-trial release and pre-trial misconduct conditional on release by defendant and judge race in Miami. The means are calculated using the Miami sample reported in Table 1. See text for additional details.

Online Appendix Table A9

p-values from Tests of Relative Racial Prejudice

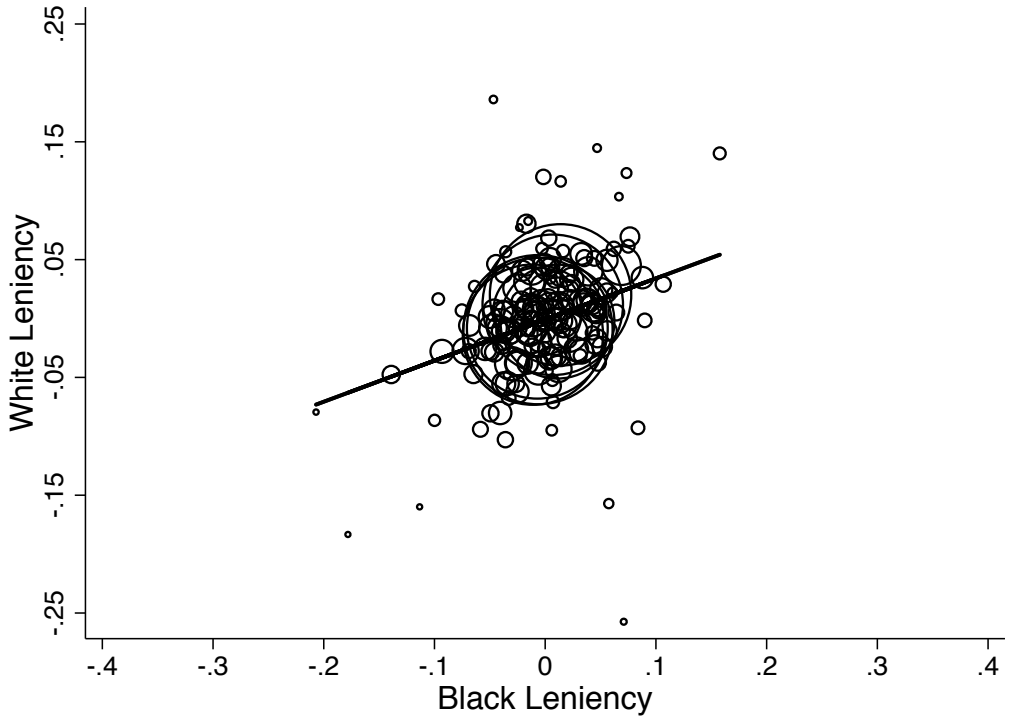
	p-Value
	(1)
Pre-Trial Release	0.782
Pre-Trial Rearrest	0.580

Notes. This table replicates the Anwar and Fang (2006) test for pre-trial release rates and pre-trial misconduct rates. This table presents bootstrapped p-values testing for relative racial bias. The null hypothesis is rejected if white judges are more lenient on white defendants, and black judges are more lenient on black defendants.

Online Appendix Table A10
Representativeness Statistics

	$\mathbb{E}(x\|Black)/\mathbb{E}(x\|White)$
Panel A: Defendant Characteristics	(1)
Male	1.026
Age at Bail Decision	0.978
Prior Offense in Past Year	1.072
Arrested on Bail in Past Year	1.048
Failed to Appear in Court in Past Year	1.028
Panel B: Charge Characteristics	
Number of Offenses	1.200
Felony Offense	1.160
Misdemeanor Only	0.866
Any Drug Offense	1.077
Any DUI Offense	0.839
Any Violent Offense	1.260
Any Property Offense	0.983
Panel C: Outcomes	
Rearrest Prior to Disposition	1.061
Drug Crime	1.059
Property Crime	1.044
Violent Crime	1.496
Failure to Appear in Court	0.983
Failure to Appear in Court or Rearrested	1.102
Observations	256,253

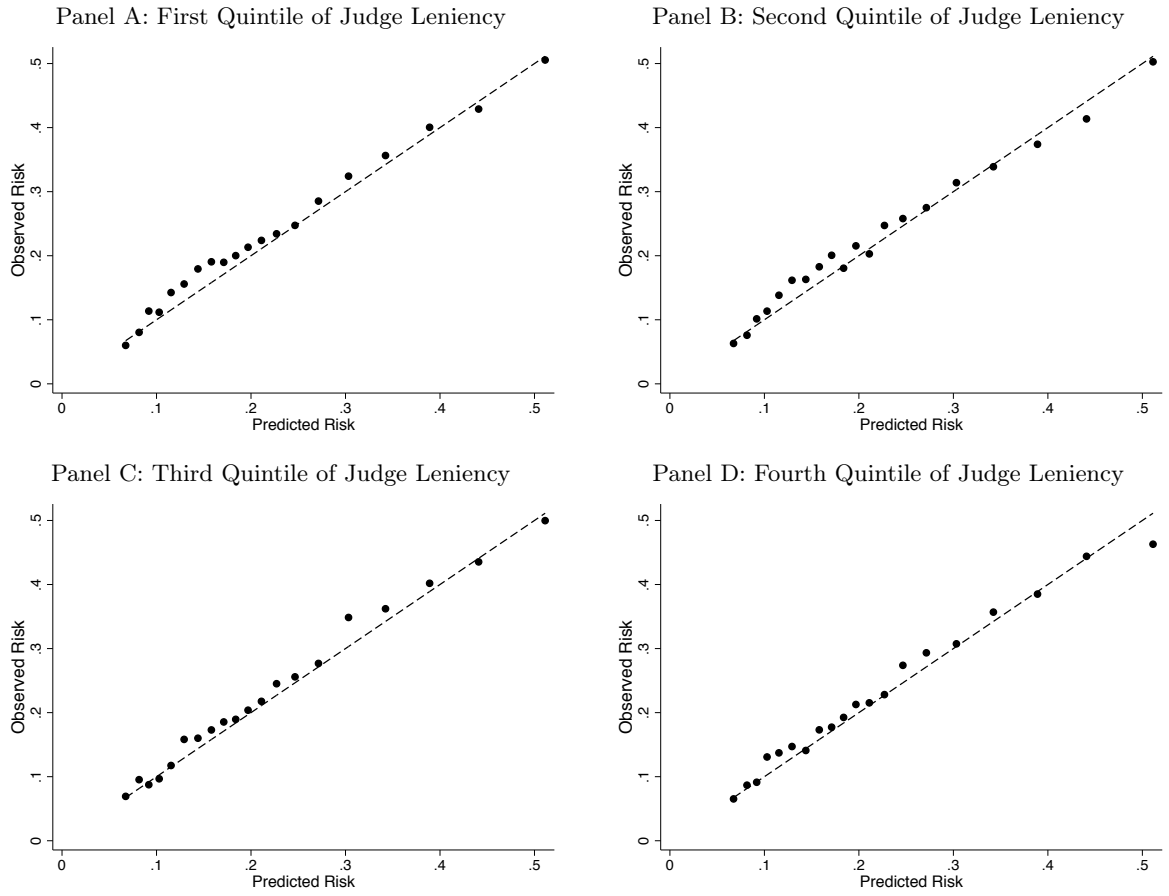
Notes. This table reports the mean of the variable listed in the row given the defendant is black, divided by the mean of the variable listed in the row given the defendant is white. The sample is described in the notes to Table 1.



Online Appendix Figure A1

Judge Leniency by Race

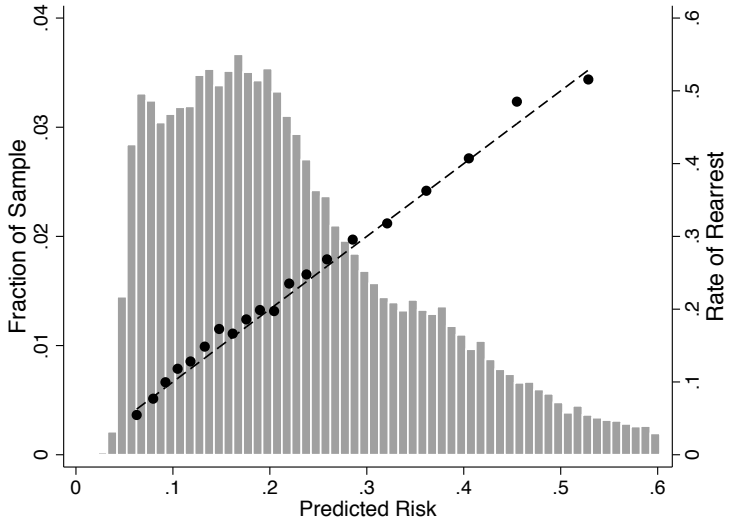
This figure show the correlation between our residualized measure of judge leniency by defendant race over all available years of data. We also plot the linear best fit line estimated using OLS.



Online Appendix Figure A2

Predicted and Actual Risk by Judge Leniency

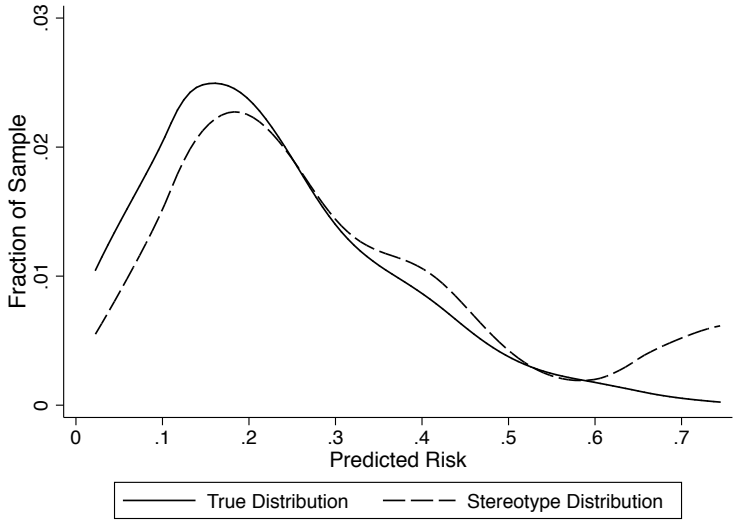
These figures plot predicted pre-trial misconduct risk against actual pre-trial misconduct for different judge-leniency quintiles. Predicted risk is calculated using only cases from the most lenient quintile of judges and the machine learning algorithm described in Online Appendix F. See the text for additional details.



Online Appendix Figure A3

Relationship between Predicted Risk and True Risk

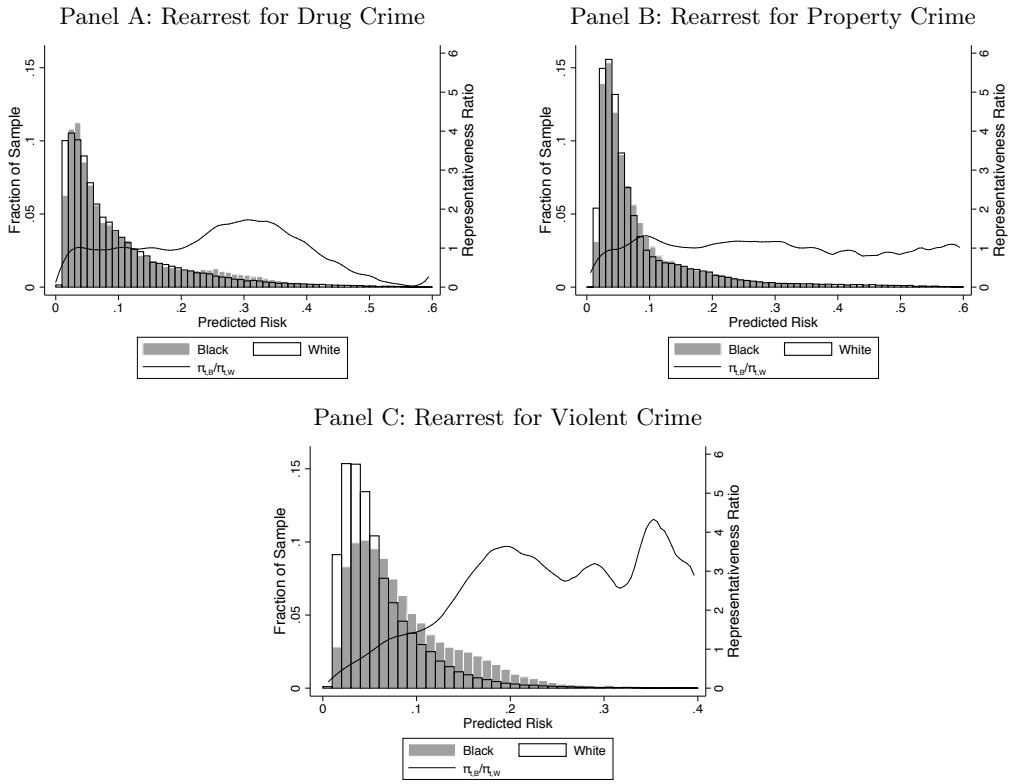
This figure reports the distribution of the pre-trial misconduct risk and plots the predicted pre-trial misconduct risk against actual pre-trial misconduct for the test sample. Predicted risk is calculated using the machine learning algorithm described in Online Appendix F. The dashed line is the 45 degree line. See the text for additional details.



Online Appendix Figure A4

Stereotyped and True Distribution of Risk for Black Defendants

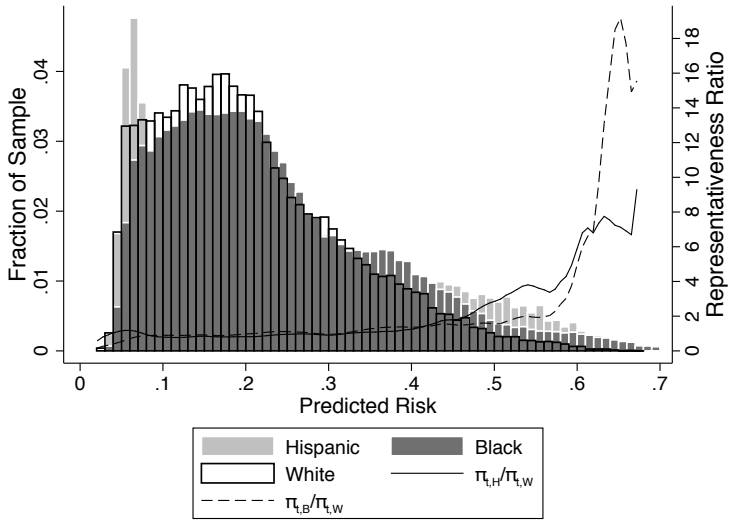
This figure plots the true distribution of risk for black defendants alongside the perceived distribution of risk for black defendants. The stereotyped beliefs are generated by a representativeness-based discounting model with $\theta = 1.9$. This value of θ rationalizes an average release rate of black defendants equal to 68.8 percent, the actual rate of release in the data. See the text and Online Appendix F for additional details.



Online Appendix Figure A5

Crime-Specific Predicted Risk Distributions by Race

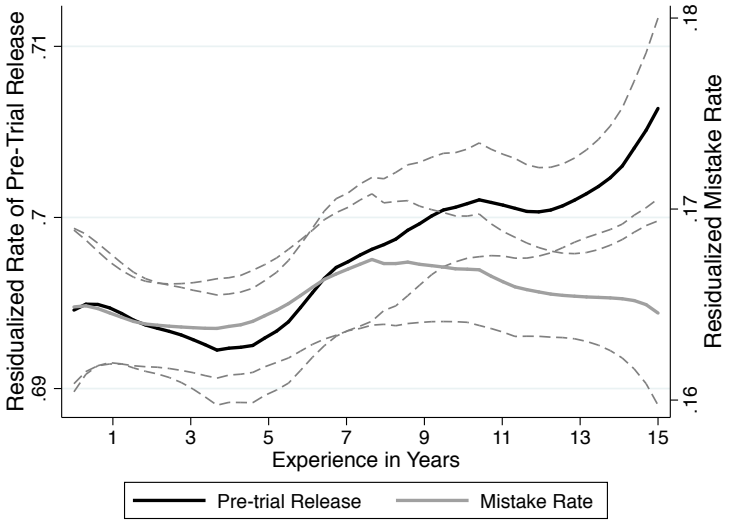
These figures report the distribution of crime-specific risk separately by defendant race. Predicted risk is calculated using the machine learning algorithm described in Online Appendix F. The solid line in each figure represents the representativeness ratio for black versus white defendants. See the text for additional details.



Online Appendix Figure A6

Predicted Risk Distribution by Hispanic and Black versus White

This figure reports the distribution of the risk of pre-trial misconduct separately by Hispanic, black, and white defendants. Predicted risk is calculated using the machine learning algorithm described in Online Appendix F. The dashed line represents the representativeness ratio for black versus white defendants and the solid line represents the representativeness ratio for Hispanic versus white defendants. See the text for additional details.



Online Appendix Figure A7

Probability of Release and Pre-trial Misconduct with Experience

This figure plots the relationship between judicial experience and both the residualized rate of pre-trial release and the residualized rate of pre-trial crime conditional on release (i.e. the mistake rate). Pre-trial release and pre-trial rearrest are both residualized using the full set of court-by-time fixed effects. See the text for additional details.

Online Appendix B: Proofs of Consistency for IV and MTE Estimators

This appendix reviews our empirical test for racial bias before providing additional details and proofs for both our IV and MTE estimation approaches. For completeness, we also include all relevant information from the main text in this appendix.

B.1. Overview

Recall that the goal of our analysis is to empirically test for racial bias in bail setting using the rate of pre-trial misconduct for white defendants and black defendants at the margin of release. Let the true weighted average of treatment effects for defendants of race r at the margin of release for judge j, α_r^j, for some weighting scheme, w^j, across all bail judges, $j = 1...J$, be given by:

$$\begin{aligned} \alpha_r^{*,w} &= \sum_{j=1}^{J} w^j \alpha_r^j \\ &= \sum_{j=1}^{J} w^j t_r^{j*} \end{aligned} \tag{18}$$

where w^j are non-negative weights which sum to one that will be discussed in further detail below. Recall that, by definition, $\alpha_r^j = t_r^{j*}$. Intuitively, $\alpha_r^{*,w}$ represents a weighted average of the treatment effects for defendants of race r at the margin of release across all judges.

Following this notation, the true average level of racial bias among bail judges, $D^{*,w}$, for the weighting scheme w^j is given by:

$$\begin{aligned} D^{*,w} &= \sum_{j=1}^{J} w^j \left(t_W^{j*} - t_B^{j*} \right) \\ &= \sum_{j=1}^{J} w^j t_W^{j*} - \sum_{j=1}^{J} w^j t_B^{j*} \\ &= \alpha_W^{*,w} - \alpha_B^{*,w} \end{aligned} \tag{19}$$

From Equation (18), we can express $D^{*,w}$ as a weighted average across all judges of the difference in treatment effects for white defendants at the margin of release and black defendants at the margin of release.

We develop two estimators for racial bias that use variation in the release tendencies of quasi-randomly assigned bail judges to identify differences in pre-trial misconduct rates at the margin of release. In theory, an estimator for $D^{*,w}$ should satisfy three criteria: (1) rely on minimal auxiliary assumptions to estimate judge-specific thresholds of release, t_r^{j*}, (2) yield statistically precise estimates of the average level of bias, $D^{*,w}$, and (3) use a policy-relevant weighting scheme, w^j. In practice, however, no single estimator can accomplish all three criteria in our setting. The two-stage least squares IV estimator, for example, relies on relatively few auxiliary assumptions

and provides statistically precise estimates by giving greater weight to more precise LATEs, but the particular weighting of the pairwise LATEs may not always yield a policy-relevant estimate of racial bias. In contrast, a fully non-parametric approach where one reports each pairwise LATE separately and allows a researcher to choose a weighting scheme can yield a policy-relevant interpretation of racial bias with minimal assumptions, but often comes at the cost of statistical precision since any particular LATE is often estimated with considerable noise. The MTE framework developed by Heckman and Vytlacil (1999, 2005) provides a third option, allowing a researcher to estimate judge-specific treatment effects for white and black defendants at the margin of release and thus choose a weighting scheme, but with estimation of racial bias for each judge, and relatedly statistical precision, coming at the cost of additional auxiliary assumptions.

In this Online Appendix, we show that both IV and MTE estimators yield qualitatively similar estimates of the average level of racial bias in our setting, suggesting that neither the choice of IV weights nor the additional parametric assumptions required under our MTE approach greatly affect our estimates. In contrast, we show that the fully non-parametric approach yields uninformative estimates of the average level of racial bias due to very imprecise estimates of the individual pairwise LATEs.

B.2. Instrumental Variables Framework

Our first estimator uses IV weights, defined as $w^j = \lambda^j$, when estimating the weighted average level of bias, $D^{*,w}$. Recall that λ^j are the standard IV weights defined in Imbens and Angrist (1994). Our IV estimator allows us to estimate a weighted average of racial bias across bail judges with relatively few auxiliary assumptions, but with the caveats that we cannot estimate judge-specific treatment effects and the weighting scheme underlying the IV estimator may not be policy relevant. If the IV weights are uncorrelated with the level of racial bias for a given judge, then our IV estimator will estimate the average level of discrimination across all bail judges. If the IV weights are correlated with the level of racial bias, however, then our IV estimator may under or overestimate the average level of racial bias across all bail judges, but may still be of policy relevance depending on the parameter of interest (e.g., an estimate of racial bias that puts more weight on judges with higher caseloads).

In this subsection, we present a formal definition of the IV-weighted level of racial bias and our IV estimator, provide proofs for consistency, discuss tests of the identifying assumptions, the interpretation of the IV-weighted estimate, and the potential bias of our IV estimator from using a discrete instrument. We then consider a re-weighting procedure that accounts for judge bias on observable non-race characteristics.

1. Definition and Consistency of IV Estimator Let the IV-weighted level of racial bias, $D^{*,IV}$ be defined as:

$$D^{*,IV} = \sum_{j=1}^{J} w^j \left(t_W^{j*} - t_B^{j*} \right) = \sum_{j=1}^{J} \lambda^j \left(t_W^{j*} - t_B^{j*} \right) \tag{20}$$

where $w^j = \lambda^j$, the instrumental variable weights defined in Imbens and Angrist (1994) and described in the main text.

Following the definition in the main text, let our IV estimator be defined as:

$$D^{IV} = \alpha_W^{IV} - \alpha_B^{IV} = \sum_{j=1}^{J} \lambda_W^j \alpha_W^{j,j-1} - \sum_{j=1}^{J} \lambda_B^j \alpha_B^{j,j-1} \tag{21}$$

where each pairwise LATE, $\alpha_r^{j,j-1}$, is again the average treatment effect of compliers between judges $j-1$ and j.

Building on the standard IV framework, we now establish the two conditions under which our IV estimator for racial bias D^{IV} provides a consistent estimate of the IV-weighted level of racial bias, $D^{*,IV}$.

First Condition for Consistency: The first condition for our IV estimator D^{IV} to provide a consistent estimate of $D^{*,IV}$ is that our judge leniency measure Z_i is continuously distributed over some interval $[\underline{z}, \bar{z}]$. Formally, as our instrument becomes continuous, for any judge j and any $\epsilon > 0$, there exists a judge k such that $|z_j - z_k| < \epsilon$.

PROPOSITION B.1. As Z_i becomes continuously distributed, each race-specific IV estimate, α_r^{IV}, converges to a weighted average of treatment effects for defendants at the margin of release.

Proof of Proposition B.1. To see why this proposition holds, first define the treatment effect for a defendant at the margin of release at z_j as:

$$\alpha_r^j = \alpha_r(z = z_j) = \lim_{dz \to 0} \mathbb{E}[Y_i(1) - Y_i(0) | Released_i(z) - Released_i(z - dz) = 1] \tag{22}$$

With a continuous instrument Z_i, Angrist, Graddy, and Imbens (2000) show that the IV estimate, α_r^{IV}, converges to:

$$\alpha_r = \int \lambda_r(z) \alpha_r(z) dz \tag{23}$$

where the weights, $\lambda_r(z)$ are given by:

$$\lambda_r(z) = \frac{\frac{\partial Released_r}{\partial z}(z) \cdot \int_z^{\bar{z}} (y - \mathbb{E}[z]) \cdot f_r^z(y) dy}{\int_{\underline{z}}^{\bar{z}} \frac{\partial Released_r}{\partial z}(v) \cdot \int_v^{\bar{z}} (y - \mathbb{E}[z]) \cdot f_r^z(y) dy dv} \tag{24}$$

where $\frac{\partial Released_r}{\partial z}$ is the derivative of the probability of release with respect to leniency and f_r^z is the probability density function of leniency. If $\frac{\partial Released_r}{\partial z} \geq 0$ for all z, then the weights are nonnegative. Therefore, as Z_i becomes continuously distributed, our race-specific IV estimate will return a weighted average of treatment effects of defendants on the margin of release. ■

Second Condition for Consistency: The second condition for our IV estimator D^{IV} to provide a consistent estimate of $D^{*,IV}$ is that the weights λ_r^j must be equal across race. Equal weights ensure that the race-specific IV estimates from Equation (7) in the main text, α_W^{IV} and α_B^{IV}, provide the same weighted averages of $\alpha_W^{j,j-1}$ and $\alpha_B^{j,j-1}$. If the weights $\lambda_W^j = \lambda_B^j = \lambda^j$, our IV estimator can then be rewritten as a simple weighted average of the difference in pairwise LATEs for white and black defendants:

$$D^{IV} = \sum_{j=1}^{J} \lambda^j (\alpha_W^{j,j-1} - \alpha_B^{j,j-1}) \tag{25}$$

Proof of Consistency. We combine these two conditions to establish the consistency of our IV estimator. Recall that our IV estimator D^{IV} provides a consistent estimate of racial bias $D^{*,IV}$ if (1) Z_i is continuous and (2) λ_r^j is constant by race.

To begin, we write D^{IV} as:

$$\begin{aligned} D^{IV} &= \alpha_W^{IV} - \alpha_B^{IV} \\ &= \sum_{j=1}^{J} \lambda_W^j \alpha_W^{j,j-1} - \sum_{j=1}^{J} \lambda_B^j \alpha_B^{j,j-1} \end{aligned} \tag{26}$$

If $\lambda_r^j = \lambda^j$, then:

$$D^{IV} = \sum_{j=1}^{J} \lambda^j \left(\alpha_W^{j,j-1} - \alpha_B^{j,j-1} \right) \tag{27}$$

Following Proposition B.1, as Z_i becomes continuously distributed, we can rewrite D^{IV} as:

$$\begin{aligned} D^{IV} &= \int \lambda(z) \left(\alpha_W(z) - \alpha_B(z) \right) dz \\ &= D^{*,IV} \end{aligned} \tag{28}$$

Therefore, in the limit, D^{IV} estimates a weighted average of differences in treatment effects for defendants at the margin of release, and therefore provides a consistent estimate of $D^{*,IV}$. ■

2. Empirical Implementation

a. Testing the equal weights assumption: A key assumption for the consistency of our IV estimator is that the IV weights are the same across race. Following Cornelissen et al. (2016), we calculate white and black IV weights for each judge-by-year cell by replacing the terms in Equation (24) with their sample analogues. Noting that our instrument is linear by construction and, as a result, that $\frac{\partial Released_r(z)}{\partial z}(z) = c$, we drop the term $\frac{\partial Released_r(z)}{\partial z}(z) = c$, as this appears in both the numerator and denominator of Equation (24). We then use kernel density methods to retrieve an estimate $\hat{f}_r^z$, which is the density of leniency for race r. With this estimate of the density of leniency for race r, we can plug in the sample analogue of $\mathbb{E}[z]$ and use numerical integration to estimate the remaining terms and estimate IV weights by race for each point in the distribution.

One implication of the equal weights assumption is that the distributions of black and white IV weights over the distribution of judge leniency are statistically identical. To implement this test, Online Appendix Figure B1 plots the IV weights for each judge-by-year cell, the level of our variation, by race. The distributions of black and white IV weights are visually indistinguishable from each other and a Kolmogorov–Smirnov test cannot reject the hypothesis that the two estimated distributions are drawn from the same continuous distribution ($p = .431$).

A second implication of the equal weights assumption is that the relationship between the black IV weights and the white IV weights should fit a 45-degree line up to sampling error. Online Appendix Figure B2 plots the black IV weights and the white IV weights for each judge-by-year cell, where we discretize the continuous weights to retrieve an estimate of the weights for each judge-by-year cell and then normalize the weights so that the weights sum to one (in the continuous version the weights integrate to one). The black and white IV weights for each judge-by-year cell are highly correlated across race. To formally test for violations of the equal weights assumption, we regress each black IV weight for each judge-by-year cell on the white IV weight for the same cell. This regression yields a coefficient on the white IV weight equal to 1.028 with a standard error of 0.033. Thus, both tests suggest that our assumption of equal IV weights by race is satisfied in the data.

b. Understanding the IV weights: We now investigate the relationship between IV weights and judge characteristics to better understand the economic interpretation of an IV-weighted estimate of racial bias. Online Appendix Table B1 presents OLS estimates of IV weights in each judge-by-year cell on observable judge-by-year characteristics separately by race. The correlation between the IV weights and both average leniency and whether the judge is a minority is statistically zero in both the white and black distribution, with only a weak correlation between the IV weights and judge experience in a given year. Conversely, the IV weights are positively correlated with the number of cases in a judge-by-year cell and a judge being from Philadelphia (where each judge-by-year cell has more observations). These results suggest that the additional precision in our IV regressions comes, at least in part, from placing more weight on judge-by-year cells with more observations. The IV weights are also positively correlated with judge-by-year specific estimates of racial bias

(estimated using the MTE approach discussed in Section B.4 below), although not differentially by defendant race. The positive correlation between the IV weights and the judge-by-year estimates of bias implies that the IV-weighted estimate of racial bias will be larger than an equal-weighted estimate of racial bias. All of our IV results should be interpreted with these correlations in mind.

c. Bounding the maximum bias of the IV estimator with a discrete instrument: Our approach assumes continuity of the instrument Z_i. If the instrument is discrete, we can characterize the maximum potential bias of our IV estimator D^{IV} relative to $D^{*,IV}$, e.g. "infra-marginality bias."

PROPOSITION B.2. If the instrumental variable weights are equal by race, the maximum bias of our IV estimator D^{IV} from $D^{*,IV}$ is given by $\max_j(\lambda^j)(\alpha^{max} - \alpha^{min})$, where α^{max} is the largest treatment effect among compliers, α^{min} is the smallest treatment effect among compliers, and λ^j is given by:

$$\lambda^j = \frac{(z_j - z_{j-1}) \cdot \sum_{l=j}^{J} \pi^l (z_l - \mathbb{E}[Z])}{\sum_{m=1}^{J} (z_j - z_{j-1}) \cdot \sum_{l=m}^{J} \pi^l (z_l - \mathbb{E}[Z])} \tag{29}$$

where π^l is the probability of being assigned to judge j.

Proof of Proposition B.2. To prove that this proposition holds, we proceed in five steps. First, we show that $D^{*,IV}$ is equal to D^{IV} plus a bias term, which we refer to as "infra-marginality bias." Second, we derive an upper bound for the bias term by replacing $\alpha_W^{j,j-1}$ with its minimum possible value for every judge j, and we derive a lower bound by replacing $\alpha_B^{j,j-1}$ with its maximum value for every j. Third, we show that the upper bound and lower bound of D^{IV} both converge to $D^{*,IV}$ as Z_i becomes continuously distributed. Fourth, we develop a formula for the maximum potential bias with a discrete instrument using the derived upper and lower bounds, and provide intuition for how we derive this estimation bias. Fifth, we show how to empirically estimate the maximum potential bias in the case of a discrete instrument.

Recall that under our theory model, compliers for judge j and $j-1$ are individuals such that $t_r^{j-1*} < \mathbb{E}[\alpha_i | r_i] \leq t_r^{j*}$. Under this definition of compliers, we know that:

$$\alpha_r^{j,j-1} \in (t_r^{j-1*}, t_r^{j*}] \tag{30}$$

Note that we can rewrite $D^{*,IV}$ as:

$$\begin{aligned} D^{*,IV} &= \sum_{j=1}^{J} \lambda^j \left(t_W^{j*} - t_B^{j*} \right) \\ &= \sum_{j=1}^{J} \lambda^j \left(\alpha_W^{j,j-1} - \alpha_B^{j,j-1} \right) + \sum_{j=1}^{J} \lambda^j \left(t_W^{j*} - \alpha_W^{j,j-1} \right) + \sum_{j=1}^{J} \lambda^j \left(\alpha_B^{j,j-1} - t_B^{j*} \right) \\ &= D^{IV} + \underbrace{\sum_{j=1}^{J} \lambda^j \left(t_W^{j*} - \alpha_W^{j,j-1} \right) + \sum_{j=1}^{J} \lambda^j \left(\alpha_B^{j,j-1} - t_B^{j*} \right)}_{\text{infra-marginality bias}} \end{aligned} \tag{31}$$

The second line follows from adding and subtracting $\sum_{j=1}^{J} \lambda^j \alpha_W^{j,j-1}$ and $\sum_{j=1}^{J} \lambda^j \alpha_B^{j,j-1}$ to $D^{*,IV}$ and rearranging terms. The third line follows from assuming equal IV weights by race. Equation (31) shows that $D^{*,IV}$ is equal to D^{IV} plus a bias term, which we refer to as "infra-marginality bias."

We will now derive an upper bound for $D^{*,IV}$. First, note that Equation (30) implies $\alpha_B^{j,j-1} \leq t_B^{j*}$. Therefore $\sum_{j=1}^{J} \lambda^j \left(\alpha_B^{j,j-1} - t_B^{j*} \right) \leq 0$, given $\lambda^j \geq 0$ for all j. We can drop this term from Equation (31) to obtain an upper bound on $D^{*,IV}$:

$$\begin{aligned} D^{*,IV} &\leq D^{IV} + \sum_{j=1}^{J} \lambda^j \left(t_W^{j*} - \alpha_W^{j,j-1} \right) \\ &< D^{IV} + \sum_{j=1}^{J} \lambda^j \left(t_W^{j*} - t_W^{j-1*} \right) \end{aligned} \tag{32}$$

where the second line follows from Equation (30) ($t_W^{j-1*} < \alpha_W^{j,j-1}$).

Using similar logic, we can also derive a lower bound for $D^{*,IV}$. Equation (30) implies $t_W^{j*} \geq \alpha_W^{j,j-1}$. Therefore $\sum_{j=1}^{J} \lambda^j \left(t_W^{j*} - \alpha_W^{j,j-1} \right) \geq 0$, given $\lambda^j \geq 0$ for all j. We can drop this term from Equation (31) to obtain a lower bound on $D^{*,IV}$:

$$\begin{aligned} D^{*,IV} &\geq D^{IV} + \sum_{j=1}^{J} \lambda^j \left(\alpha_B^{j,j-1} - t_B^{j*} \right) \\ &= D^{IV} - \sum_{j=1}^{J} \lambda^j \left(t_B^{j*} - \alpha_B^{j,j-1} \right) \\ &> D^{IV} - \sum_{j=1}^{J} \lambda^j \left(t_B^{j*} - t_B^{j-1*} \right) \end{aligned} \tag{33}$$

where again, the last line follows from Equation (30) ($t_B^{j-1*} < \alpha_B^{j,j-1}$).

We can now bound $D^{*,IV}$ using Equation (33) and Equation (32):

$$D^{IV} - \sum_{j=1}^{J} \lambda^j \left(t_B^{j*} - t_B^{j-1*} \right) < D^{*,IV} < D^{IV} + \sum_{j=1}^{J} \lambda^j \left(t_W^{j*} - t_W^{j-1*} \right) \tag{34}$$

It is straightforward to see that the infra-marginality bias goes to zero as Z_i becomes continuous. Given that λ^j are non-negative weights which sum to one, $\sum_{j=1}^{J} \lambda^j \left(t_r^{j*} - t_r^{j-1*} \right) \leq \max_j (t_r^{j*} - t_r^{j-1*})$ (i.e. the average is less than the maximum). Therefore, if Z_i becomes continuous, then $t_r^{j*} - t_r^{j-1*} \to 0$ for all j, and so infra-marginality bias shrinks to zero. Intuitively, at the limit, every complier is at the margin, and so there is no infra-marginality bias. As a result, D^{IV} converges to $D^{*,IV}$ as Z_i becomes continuous.

Note that $t_r^{j*} - t_r^{j-1*}$ is positive for all j, implying $\sum_{j=1}^{J} \lambda^j \left(t_r^{j*} - t_r^{j-1*} \right) \leq \max_j(\lambda^j) \sum_{j=1}^{J} \left(t_r^{j*} - t_r^{j-1*} \right)$, where $\max_j(\lambda^j)$ is the maximum weight across all judges. Given the recursive structure of $\sum_{j=1}^{J} \left(t_r^{j*} - t_r^{j-1*} \right)$:

$$\max_j(\lambda^j) \sum_{j=1}^{J} \left(t_r^{j*} - t_r^{j-1*} \right) = \max_j(\lambda^j)(t_r^{J*} - t_r^{0*}) \tag{35}$$

Note that $t_r^{J*} = \alpha_r^{max}$ (i.e. the largest treatment effect is associated with the most lenient judge) and $t_r^{0*} = \alpha_r^{min}$ (i.e. the smallest treatment effect is associated with the most strict judge). Therefore, letting α^{max} and α^{min} equal the maximum treatment effect and minimum treatment effect respectively across races, yields:

$$D^{IV} - \max_j(\lambda^j)(\alpha^{max} - \alpha^{min}) < D^{*,IV} < D^{IV} + \max_j(\lambda^j)(\alpha^{max} - \alpha^{min}) \tag{36}$$

which proves Proposition B.2. In other words, the maximum bias of our IV estimator D^{IV} from $D^{*,IV}$ is given by $\max_j(\lambda^j)(\alpha^{max} - \alpha^{min})$. ■

Next, we simplify these bounds to retrieve estimable bounds. Note that $\alpha^{max} \leq 1$ and $\alpha^{min} \geq 0$ in theory, which implies $(\alpha^{max} - \alpha^{min}) \leq 1$. Therefore, the bounds in Equation (36) can be re-written as:

$$D^{IV} - \max_j(\lambda^j) < D^{*,IV} < D^{IV} + \max_j(\lambda^j) \tag{37}$$

Rearranging terms yields:

$$-\max_j(\lambda^j) < D^{*,IV} - D^{IV} < \max_j(\lambda^j) \tag{38}$$

Under this worst-case assumption, the maximum bias of our IV estimator D^{IV} from $D^{*,IV}$ is given by $\max_j(\lambda^j)$.

To understand the intuition of our maximum bias formula, note that under Proposition B.2,

the maximum bias of D^{IV} relative to $D^{*,IV}$ decreases as (1) the heterogeneity in treatment effects among compliers decreases ($\alpha^{max} \rightarrow \alpha^{min}$) and (2) the maximum of the judge weights decreases ($\max_j(\lambda^j) \rightarrow 0$), as would occur when there are more judges distributed over the range of the instrument. If treatment effects are homogeneous among compliers such that $\alpha^{max} = \alpha^{min}$, our IV estimator D^{IV} continues to provide a consistent estimate of $D^{*,IV}$. In practice, we calculate the maximum bias of our estimator under the worst-case assumption of treatment effect heterogeneity (i.e. $\alpha^{max} - \alpha^{min} = 1$, the maximum possible value). Because the weights λ^j are identified in our data, the maximum bias due to infra-marginality concerns can be conservatively estimated to be equal to $\max_j(\lambda^j)$.

In general, the IV weights, λ^j, will not be equal across judges. In particular, the weights depend partially on the share of compliers between any two adjacent judges. For example, if there are more infra-marginal defendants for lenient judges, then lenient judges will be given more weight in the estimation of racial bias. However, our bounding procedure of the maximum bias does not rely on any assumption about equal weights across judges. For example, consider an extreme case where although there are many judges, defendants are only infra-marginal to the most-strict and second most-strict judge. Then, the entire share of compliers will be defendants who are detained by the most-strict judge and released by the second most-strict judge. Therefore, the pairwise LATE for the most-strict judge and the second most-strict judge will receive the entire weight in estimating the effect of release on the probability of pre-trial misconduct. In this case, we would conclude that the maximum bias of our estimator is equal to one, and therefore, we would be unable to provide informative bounds on the true level of racial bias.

We can illustrate this point with a simple two judge case where both judges use the same release thresholds for both white and black defendants, $t_W^{j*} = t_B^{j*}$, such that there is no racial bias, $D^{*,IV} = 0$. Suppose that the more lenient judge releases defendants with an expected pre-trial misconduct rate of less than 20 percent, while the more strict judge releases defendants with an expected pre-trial misconduct rate of less than 10 percent. Then, the race-specific LATEs estimated using our IV strategy are the average treatment effects of all defendants with expected misconduct rates between 10 and 20 percent. Within this range of compliers, suppose that all black defendants have expected rates of pre-trial misconduct of 10 percent, while all white defendants have expected rates of pre-trial misconduct of 20 percent. Then, our IV estimator will yield a LATE for whites ($\alpha_W^{IV} = 0.2$) that is larger in magnitude than the LATE for blacks ($\alpha_B^{IV} = 0.1$), causing us to estimate $D^{IV} = 0.1 > 0$. Our IV estimator would thus lead us to incorrectly conclude that there was racial bias. A similar exercise can be used to show that we may find $D^{IV} = 0$ even if $D^{*,IV} > 0$. Under the worst-case scenario where we assume the maximum heterogeneity in treatment effects ($\alpha^{max} - \alpha^{min} = 1$), the maximum infra-marginality bias is $\max_j(\lambda^j) = 1$ because 100 percent of compliers fall within the two judges. In this case, infra-marginality bias makes our IV estimator uninformative on the true level of racial bias. However, using the same logic, it is straightforward to show that the magnitude of this infra-marginality bias decreases when there are many judges because the share of compliers within any two judges decreases, thus decreasing $\max_j(\lambda^j)$.

We can now illustrate how we empirically estimate the maximum potential bias of our IV estimator from $D^{*,IV}$ using the formula in Proposition B.2. Again, because we do not observe $\alpha^{max} - \alpha^{min}$, we take the most conservative approach and assume that this value is equal to 1. Imbens and Angrist (1994) show that the instrumental variables weights, λ^j, for a discrete multi-valued instrument are given by the following formula:

$$\lambda^j = \frac{(Pr(Released|z_j) - Pr(Released|z_{j-1})) \cdot \sum_{l=j}^{J} \pi^l (g(z_l) - \mathbb{E}[g(Z)])}{\sum_{m=1}^{J} (Pr(Released|z_m) - Pr(Released|z_{m-1})) \cdot \sum_{l=m}^{J} \pi^l (g(z_l) - \mathbb{E}[g(Z)])} \tag{39}$$

where π^l is the probability a defendant is assigned to judge l, $g(z_l)$ is a function of the instrument, and $Pr(Released|z_j)$ if the probability a defendant is released if assigned to judge j. While λ^j is not indexed by r, we estimate the weights completely separately by race. To proceed, we residualize both the endogenous variable *Released* and the judge leniency instrument using all exogenous regressors. An instrumental variables regression utilizing residualized variables yields a numerically identical estimate as the specification in the main text (Evdokimov and Kolesár 2018). To estimate the weight λ^j we simply replace each expression in Equation (39) with the empirical counterpart. Formally:

$$Pr(Released|z_j) - Pr(Released|z_{j-1}) = \mathbb{E}[\ddot{R}|\ddot{z}_j] - \mathbb{E}[\ddot{R}|\ddot{z}_{j-1}] \tag{40}$$

where $\ddot{R}$ is *Released* residualized by the exogenous regressors and $\ddot{z}_j$ is the residualized value of the instrument. Since we use residualized judge leniency as the instrument we replace $g(\ddot{z}_l) = \ddot{z}_l$. Lastly, we replace π^j and $\mathbb{E}[Z]$ with their empirical counterparts:

$$\hat{\pi}^j = \sum_{i=1}^{N} \frac{\mathbb{1}\{\ddot{Z}_i = \ddot{z}_j\}}{N} \tag{41}$$

$$\mathbb{E}[Z] = \frac{1}{N} \sum_{i=1}^{N} \ddot{Z}_i \tag{42}$$

Plugging these quantities into the formula for the weights yields an estimate of the weight attached to each pairwise LATE. We then take the maximum of our weights and interpret this estimate as the maximum potential bias between our IV estimator and $D^{*,IV}$. This procedure yields a maximum bias of 0.011 or 1.1 percentage points.

From Equation (37), we know:

$$D^{*,IV} < D^{IV} + \max_j(\lambda^j) = D^{IV} + 0.011$$
$$D^{*,IV} > D^{IV} - \max_j(\lambda^j) = D^{IV} - 0.011$$

Therefore, in our setting, $D^{*,IV}$ is bounded within 1.1 percentage points of our IV estimate for

racial bias. ■

3. Re-weighting Procedure to Allow Judge Preferences for Non-Race Characteristics In this subsection, we show that a re-weighting procedure using our IV estimator can be used to estimate direct racial bias (i.e. racial bias which cannot be explained by the composition of crimes). To begin, let the weights for all white defendants be equal to 1. We construct the weights for a black defendant with observables equal to $\mathbf{X}_i = x$ as:

$$\Psi(x) = \frac{Pr(W|x)Pr(B)}{Pr(B|x)Pr(W)} \tag{43}$$

where $Pr(W|x)$ is the probability of being white given observables $\mathbf{X}_i = x$, $Pr(B|x)$ is the probability of being black given observables $\mathbf{X}_i = x$, $Pr(B)$ is the unconditional probability of being black, and $Pr(W)$ is the unconditional probability of being white.

Define the covariate-specific LATE as:

$$\alpha_r^{j,j-1}(x) = \mathbb{E}[Y_i(1) - Y_i(0)|R_i(z_j) - R_i(z_{j-1}) = 1|r_i = r, \mathbf{X}_i = x] \tag{44}$$

As noted by Fröhlich (2007) and discussed in Angrist and Fernández-Val (2013), the unconditional LATE can be expressed as:

$$\alpha_r^{j,j-1} = \sum_{x \in X} \alpha_r^{j,j-1}(x) \frac{Pr(Released|z_j, x, r) - Pr(Released|z_{j-1}, x, r)}{Pr(Released|z_j, r) - Pr(Released|z_{j-1}, r)} P(x|r) \tag{45}$$

We assume:

$$\frac{Pr(Released|z_j, x, r) - Pr(Released|z_{j-1}, x, r)}{Pr(Released|z_j, r) - Pr(Released|z_{j-1}, r)} = \xi(x) \tag{46}$$

In words, while the first stage may vary based on covariates, it varies in the same way for white and black defendants. Therefore, in the re-weighted sample, $\alpha_B^{j,j-1}$ is given by:

$$\begin{aligned}
\alpha_B^{j,j-1} &= \sum_{x \in X} \alpha_B^{j,j-1}(x)\xi(x)Pr(x|B)\Psi(x) \\
&= \sum_{x \in X} \alpha_B^{j,j-1}(x)\xi(x)Pr(x|B)\frac{Pr(W|x)Pr(B)}{Pr(B|x)Pr(W)} \\
&= \sum_{x \in X} \alpha_B^{j,j-1}(x)\xi(x)\frac{Pr(B|x)Pr(x)}{Pr(B)}\frac{Pr(W|x)Pr(B)}{Pr(B|x)Pr(W)} \\
&= \sum_{x \in X} \alpha_B^{j,j-1}(x)\xi(x)\frac{Pr(W|x)Pr(x)}{Pr(W)} \\
&= \sum_{x \in X} \alpha_B^{j,j-1}(x)\xi(x)Pr(x|W)
\end{aligned}$$

where line 2 follows by plugging in the formula for $\Psi(x)$ and lines 3 and 5 follow from Bayes' rule. These steps closely follow DiNardo, Fortin, and Lemieux (1996), although our parameter of interest is a treatment effect rather than a distribution. Given that the weights for all white defendants are equal to 1, D^{IV} is given by:

$$D^{IV} = \sum_{j=1}^{J} \lambda^j \left(\sum_{x \in X} \xi(x) Pr(x|W) \left(\alpha_W^{j,j-1}(x) - \alpha_B^{j,j-1}(x) \right) \right) \tag{47}$$

■

B.3. Non-Parametric Pairwise LATE Framework

A second approach to estimating the average level of racial bias is to estimate each pairwise LATE separately and then impose the preferred weighting scheme across these non-parametric estimates. We consider, for example, an approach that places equal weight on each judge to estimate the average level of racial bias across judges all judges in the sample. This fully non-parametric approach can yield a policy-relevant interpretation of racial bias with minimal assumptions, but often comes at the cost of statistical precision since any particular LATE is often estimated with considerable noise.

In this subsection, we present a formal definition of the equal-weighted level of bias and our non-parametric estimator, provide proofs for consistency, and evaluate the feasibility of this non-parametric approach using Monte Carlo simulations.

1. Definition and Consistency of Pairwise LATE Estimator Let the equal-weighted LATE estimate of racial bias based on the non-parametric pairwise estimates, $D^{*,PW}$ be defined as:

$$\begin{aligned} D^{*,PW} &= \sum_{j=1}^{J} w^j \left(t_W^{j*} - t_B^{j*} \right) \\ &= \sum_{j=1}^{J} \frac{1}{J} \left(t_W^{j*} - t_B^{j*} \right) \end{aligned} \tag{48}$$

where $w^j = \frac{1}{J}$, such that $D^{*,PW}$ can be interpreted as the average level of racial bias across judges—an estimate with clear economic interpretation.

Let the equal-weighted pairwise LATE estimator of racial bias, D^{PW}, be defined as:

$$D^{PW} = \sum_{j=1}^{J} \frac{1}{J} (\alpha_W^{j,j-1} - \alpha_B^{j,j-1}) \tag{49}$$

where each pairwise LATE, $\alpha_r^{j,j-1}$, is again the average treatment effect of compliers between judges $j-1$ and j.

Conditions for Consistency: Following the proofs for the IV estimator, D^{PW} provides a consistent estimate of racial bias $D^{*,PW}$ if (1) Z_i is continuous and (2) w^j is constant by race, which is satisfied

because the weights are chosen ex post to be equal ($w^j = \frac{1}{J}$).

2. Empirical Implementation

a. Estimating the pairwise LATEs: We estimate non-parametric LATEs using the following Wald estimator for each pair of judges j and judge $j-1$:

$$\hat{\alpha}_r^{j,j-1} = \frac{\mathbb{E}[Y_i|Z_i = z_j, r] - \mathbb{E}[Y_i|Z_i = z_{j-1}, r]}{\mathbb{E}[Released_i|Z_i = z_j, r] - \mathbb{E}[Released_i|Z_i = z_{j-1}, r]} \tag{50}$$

where $\mathbb{E}[Y_i|Z_i = z_j, r]$ is the probability a defendant of race r assigned to judge j is rearrested and $\mathbb{E}[Released_i|Z_i = z_j, r]$ is the probability a defendant of race r assigned to judge j is released. Following the above discussion, our equal-weighted estimate of racial bias is equal to the simple difference between the average estimated pairwise LATE for white defendants and the average estimated pairwise LATE for black defendants.

b. Monte Carlo simulation: As discussed above, a fully non-parametric approach can yield a policy-relevant interpretation of racial bias with minimal assumptions, but often comes at the cost of statistical precision since any particular LATE is often estimated with considerable noise. We therefore begin by examining the performance of our non-parametric estimator using Monte Carlo simulations. Specifically, we create a simulated dataset with 170 judges, where each judge is assigned 500 cases with black defendants and 500 cases with white defendants. The latent risk of rearrest before disposition for each defendant is drawn from a uniform distribution between 0 and 1. Each judge releases defendants if and only if the risk of rearrest is less than his or her race-specific threshold. In the simulated data, each judge's threshold for white defendants is set to match the distribution of judge leniencies observed in the true data. For each judge, we then impose a 10 percentage point higher threshold for black defendants, so that the "true" level of racial bias in the simulated data is exactly equal to 0.100. The probability that a released defendant is rearrested ($Y_i = 1$) conditional on release is equal to the risk of the released defendant.

In each draw of the simulated data, we estimate non-parametric LATEs using the Wald estimator described above. Our estimate of racial bias in each draw of the simulated data is equal to the difference between the average release threshold for white defendants and the average release threshold for black defendants. We repeat this entire process 500 times and plot the resulting estimates of the average level of racial bias across all bail judges.

Panel A of Online Appendix Figure B3 presents the results from this Monte Carlo exercise. The average level of racial bias across all simulations is equal to 0.125, close to the true level. However, the variance of the estimates is extremely large, with nearly 20 percent of the simulations yielding an estimate of racial bias that is greater than one in absolute value. The high variance in the estimates stems from weak first stages between judges that are very close in the leniency distribution. We conclude from this exercise that a fully non-parametric approach yields uninformative estimates of average racial bias in our setting, and do not explore this approach further.[1]

[1]In unreported results, we also examine the performance of a non-parametric estimator where estimates of α_r^j

B.4. Marginal Treatment Effects Framework

Our final estimator uses the MTE framework developed by Heckman and Vytlacil (1999, 2005) to estimate the average level of bias, $D^{*,w}$, where we impose equal weights for each judge. The MTE framework allows us to estimate judge-specific treatment effects for white and black defendants at the margin of release and choose a weighting scheme across all judges, but with the identification and estimation of the judge-specific estimates, t_r^{j*}, coming at the cost of additional auxiliary assumptions.

In this subsection, we present a formal definition of the equal-weighted level of bias and our MTE estimator, provide details on the mapping of the MTE framework to our test of racial bias, provide proofs for consistency, and discuss the details of the empirical implementation and tests of the parametric assumptions.

1. Definition and Consistency of MTE Estimator

Following the discussion of the equal-weighted non-parametric estimator, let the equal-weighted MTE estimate of racial bias, $D^{*,MTE}$ be defined as:

$$D^{*,MTE} = \sum_{j=1}^{J} w^j \left(t_W^{j*} - t_B^{j*} \right) = \sum_{j=1}^{J} \frac{1}{J} \left(t_W^{j*} - t_B^{j*} \right) \tag{51}$$

where $w^j = \frac{1}{J}$, such that $D^{*,MTE}$ can again be interpreted as the average level of racial bias across judges.

Let our equal-weighted MTE estimator of racial bias, D^{MTE}, be defined as:

$$D^{MTE} = \sum_{j=1}^{J} \frac{1}{J} \left(MTE_W(p_W^j) - MTE_B(p_B^j) \right) \tag{52}$$

where p_r^j is the probability judge j releases a defendant of race r (i.e. judge j's propensity score) and $MTE_r(p_r^j)$ is the estimated MTE at the propensity score for judge j calculated separately for each defendant of race r.

2. MTE Framework:

To formally map our model of racial bias from the main text to the MTE framework developed by Heckman and Vytlacil (2005), we first characterize judge j's pre-trial release decision as:

$$Released_i(z_j, r) = \mathbb{1}\{\mathbb{E}[\alpha_i | r] \leq t_r^j(\mathbf{V}_i)\} \tag{53}$$

are formed using a Wald estimator between judge j to judge $j-k$, where $k > 1$. We find that increasing k decreases variance in the simulated estimates, but increases estimation bias, as judges further away in the distribution are used to estimate judge j's threshold. Even with relatively large k, we find the MTE procedure described in Section B.4 is more precise than the pairwise LATE procedure.

where $Released_i(z_j, r)$ indicates the probability defendant i of race r is released by judge j, and α_i, and $t_r^j(\mathbf{V}_i)$ are defined as in the main text. Vytlacil (2002) shows that under our assumptions of independence and monotonicity, the treatment decision can be written as a latent-index model of the following form:

$$Released_i(z_j, r) = \mathbb{1}\{U_{i,r} \leq p_r^j\}$$

where $U_{i,r} \in [0, 1]$ by construction. In this latent-index model, defendants with $U_{i,r} \leq p_r^j$ are released, defendants with $U_{i,r} > p_r^j$ are detained, and defendants with $U_{i,r} = p_r^j$ are on the margin of release for judge j.

Following Heckman and Vytlacil (2005), we define the race-specific marginal treatment effect as the treatment effect for defendants on the margin of release:

$$MTE_r(u) = \mathbb{E}[\alpha_i | r, U_{i,r} = u] \tag{54}$$

where $\mathbb{E}[\alpha_i | r, U_{i,r} = p_r^j]$ denotes the treatment effect for a defendant of race r who is on the margin of release to a judge with propensity score equal to p_r^j. For simplicity, we denote judge j's propensity score for defendants of race r as p_r^j.

Using the above framework, we can now describe how the race-specific MTEs defined by Equation (54) allow us estimate racial bias for each judge in our sample. First, recall that the estimand of interest is the treatment effect of pre-trial release for white and black defendants at the margin of release:

$$\alpha_r^j = \mathbb{E}[\alpha_i | r, \mathbb{E}[\alpha_i | r] = t_r^{j*}] \tag{55}$$

Because $\mathbb{E}[\alpha_i | r] = t_r^{j*}$ can be replaced with the equivalent condition, $U_{i,r} = p_r^j$, both of which state defendant i is marginal to judge j, we can equate α_r^j to the MTE function at p_r^j:

$$\begin{aligned} \alpha_r^j &= \mathbb{E}[\alpha_i | r, \mathbb{E}[\alpha_i | r] = t_r^{j*}] \\ &= \mathbb{E}[\alpha_i | r, U_{i,r} = p_r^j] \\ &= MTE_r(p_r^j) \end{aligned} \tag{56}$$

Equation (56) shows that we can use the race-specific MTEs to identify the race-specific treatment effect of each judge, α_r^j, and as a result, race-specific thresholds of release, t_r^{j*}. We can then estimate the level of racial bias for each judge j, $t_W^{j*} - t_B^{j*}$. To see this, let judge j have a propensity score to release white defendants equal to p_W^j and a propensity to release black defendants equal to p_B^j. Given Equation (56), the level of racial bias for judge j is therefore equal to $MTE_W(p_W^j) - MTE_B(p_B^j)$. From these judge-specific estimates of racial bias, we can then ex post impose equal weights across judges to estimate D^{MTE}, the average level of racial bias.

Conditions for Consistency: In addition to the assumptions required for a causal interpretation of

the IV estimator (existence, exclusion restriction, and monotonicity), our MTE estimator D^{MTE} provides a consistent estimate of $D^{*,MTE}$ if the race-specific MTEs are identified over the entire support of the propensity score calculated using variation in Z_i.

If Z_i is continuous, the local instrumental variables (LIV) estimand provides a consistent estimate of the MTE over the support of the propensity score with no additional assumptions (Heckman and Vytlacil 2005; Cornelissen et al. 2016). With a discrete instrument, however, our MTE estimator is only consistent under additional functional form restrictions that allow us to interpolate the MTEs between the values of the propensity score we observe in the data. In our MTE framework, if our specification of the MTE is flexible enough to capture the true shape of the MTE function, then there will be no infra-marginality bias. If the specification is too restrictive, then there may be misspecification bias in estimating the MTE.

Recall that our goal is to construct the average level of racial bias across judges:

$$\begin{aligned} D^{*,MTE} &= \sum_{j=1}^{J} \frac{1}{J} \left(t_W^{j*} - t_B^{j*} \right) \\ &= \sum_{j=1}^{J} \frac{1}{J} \left(\alpha_W^{j} - \alpha_B^{j} \right) \end{aligned} \tag{57}$$

With a continuous instrument, α_W^j and α_B^j are identified by evaluating $MTE(p_W^j)$ and $MTE(p_B^j)$. Heckman and Vytlacil (1999) show local instrumental variables (LIV) can be used to identify the MTE non-parametrically. With a discrete instrument, however, $MTE(p_r^j)$ is no longer non-parametrically identified.

Following Heckman and Vytlacil (2005) and Doyle (2007), we use a local polynomial function and information from the observed values of the propensity score to estimate the MTE curve over the full support of the propensity score. Specifically, we use a local quadratic estimator to approximate $\mathbb{E}[Y_i|p_r^j]$, and then estimate the MTE as the numerical derivative of the local quadratic function. In this estimation, we specify a bandwidth, and therefore use information from all judges in a given bandwidth to estimate the threshold for a given judge.

Let the estimated MTE be denoted by $\hat{MTE}(p_r^j)$. We can express our MTE estimator D^{MTE} as:

$$\begin{aligned} D^{MTE} = & \underbrace{\sum_{j=1}^{J} \frac{1}{J} \left(\hat{MTE}(p_W^j) - \hat{MTE}(p_B^j) \right)}_{\text{Estimated MTE}} + \\ & \underbrace{\sum_{j=1}^{J} \frac{1}{J} \left(MTE(p_W^j) - \hat{MTE}(p_W^j) \right) + \sum_{j=1}^{J} \frac{1}{J} \left(\hat{MTE}(p_B^j) - MTE(p_B^j) \right)}_{\text{infra-marginality bias}} \end{aligned} \tag{58}$$

In this case, infra-marginality bias arises because we allow for the possibility that the local quadratic

function does not provide enough flexibility to accurately capture the shape of the MTE. If we assume our specification of the MTE is flexible enough to capture the shape of the MTE, then $\mathbb{E}[\hat{MTE}(p_r^j)] = MTE(p_r^j)$, indicating there is no infra-marginality bias. Therefore, if we correctly specify the form of the MTE function, then D^{MTE} provides a consistent estimate of $D^{*,MTE}$:

$$
\begin{aligned}
D^{MTE} &= \sum_{j=1}^{J} \frac{1}{J} \left(MTE_W(p_W^j) - MTE_B(p_B^j) \right) \\
&= \sum_{j=1}^{J} \frac{1}{J} \left(t_W^{j*} - t_B^{j*} \right) \\
&= D^{*,MTE}
\end{aligned} \tag{59}
$$

3. Empirical Implementation

a. Estimating the MTE curve: We estimate D^{MTE} using a two-step procedure. First, we estimate the entire distribution of MTEs. To estimate each race-specific MTE, we estimate the derivative of our outcome measure (i.e. rearrest before case disposition) with respect to variation in the propensity score provided by our instrument (i.e. variation in the predicted probability of being released from the variation in judge leniency) separately for white and black defendants:

$$
MTE_W(p_W^j) = \frac{\partial}{\partial p_W^j} \mathbb{E}(\ddot{Y}_i | p_W^j, W) \tag{60}
$$

$$
MTE_B(p_B^j) = \frac{\partial}{\partial p_B^j} \mathbb{E}(\ddot{Y}_i | p_B^j, B) \tag{61}
$$

where p_r^j is the propensity score for release for judge j and defendant race r and $\ddot{Y}_i$ is rearrest residualized on all observables: an exhaustive set of court-by-time fixed effects as well as our baseline crime and defendant controls: gender, age, whether the defendant had a prior offense in the past year, whether the defendant had a prior history of pre-trial crime in the past year, whether the defendant had a prior history of failure to appear in the past year, the number of charged offenses, indicators for crime type (drug, DUI, property, violent, or other), crime severity (felony or misdemeanor), and indicators for any missing characteristics.

Following Heckman, Urzua, and Vytlacil (2006) and Doyle (2007), we begin by residualizing our measure of pre-trial misconduct, pre-trial release, and judge leniency using the full set of fixed effects and observables. We can then calculate the race-specific propensity score using a regression of the residualized release variable on our residualized measure of judge leniency, capturing only the variation in pre-trial release due to variation in the instrument.[2] Next, we compute the numerical

[2]A common approach in the MTE literature is to exploit variation in the propensity score that arises from covariates. Many treatment effect parameters, such as the average treatment effect, rely on having wide support of the propensity score. However, in practice, it is difficult to identify such strong instruments, so researchers rely on utilizing variation driven by observables. In our setting, we rely on the continuity of the propensity score to estimate the MTE, but require no assumptions concerning the range of the propensity score. In particular, the treatment

derivative of a smoothed function relating residualized pre-trial misconduct to the race-specific propensity score. Specifically, we estimate the relationship between the residualized misconduct variable and the race-specific propensity score using a local quadratic estimator. We then compute the numerical derivative of the local quadratic estimator for each race separately to obtain the race-specific MTEs. In unreported results, we also find nearly identical results using alternative estimation procedures, such as the global polynomials used in Kowalski (2016).

Second, we use the race-specific MTE distributions to calculate the level of racial bias for each judge j. We aggregate these judge-specific estimates of racial bias to calculate an equal-weighted estimate of racial bias:

$$D^{MTE} = \sum_{j=1}^{J} \frac{1}{J} \left(MTE_W(p_W^j) - MTE_B(p_B^j) \right) \tag{62}$$

We calculate standard errors of this equal-weighted estimate by bootstrapping this two-step procedure 500 times at the judge-by-shift level.

b. Monte Carlo simulation: To examine the performance of our MTE estimator, we again use a Monte Carlo simulation. Following the simulation used to test the non-parametric estimator, we create a simulated dataset with 170 judges, where each judge is assigned 500 cases with black defendants and 500 cases with white defendants. The latent risk of rearrest before disposition for each defendant is drawn from a uniform distribution between 0 and 1. Each judge releases defendants if and only if the risk of rearrest is less than his or her race-specific threshold. In the simulated data, each judge's threshold for white defendants is set to match the distribution of judge leniencies observed in the true data. For each judge, we then impose a 10 percentage point higher threshold for black defendants, so that the "true" level of racial bias in the simulated data is exactly equal to 0.100. The probability that a released defendant is rearrested ($Y_i = 1$) conditional on release is equal to the risk of the released defendant.

In each draw of the simulated data, we use the MTE estimation procedure outlined above to estimate both the race-specific MTEs and the average level of racial bias when each judge is weighted equally. We repeat this entire process 500 times and plot the resulting estimates of the average level of racial bias across all bail judges.

Panel B of Online Appendix Figure B3 presents the results from this Monte Carlo exercise. The average level of racial bias across all simulations is equal to 0.090 with a standard deviation of only 0.051. In addition, the 10^{th} percentile of estimates is equal to 0.036 and the 90^{th} percentile equal to 0.143. These results stand in sharp contrast to the statistically uninformative results from our non-parametric estimator and suggest that, in practice, our MTE estimator is likely to yield statistically precise estimates of the average level of racial bias across all bail judges.

c. Testing the MTE functional form assumption: Following Cornelissen et al. (2016), we test whether the MTE is misspecified by constructing a non-parametric IV estimate of racial bias by

effects we are interested in are identified by variation in judge leniency by definition.

taking the correct weighted average of the MTE. Specifically, we re-estimate the IV weights from Equation (24), but substitute $p(z_j)$ in for z_j, given that we estimate the MTE curve over the distribution of the propensity score, and not the distribution of leniency. We denote these weights ω_r^{IV}. As shown in Heckman and Vytlacil (2005), the IV estimate, α_r^{IV} is related to the MTE_r by:

$$\alpha_r^{IV} = \int MTE_r(u)\omega_r^{IV}(u)du \tag{63}$$

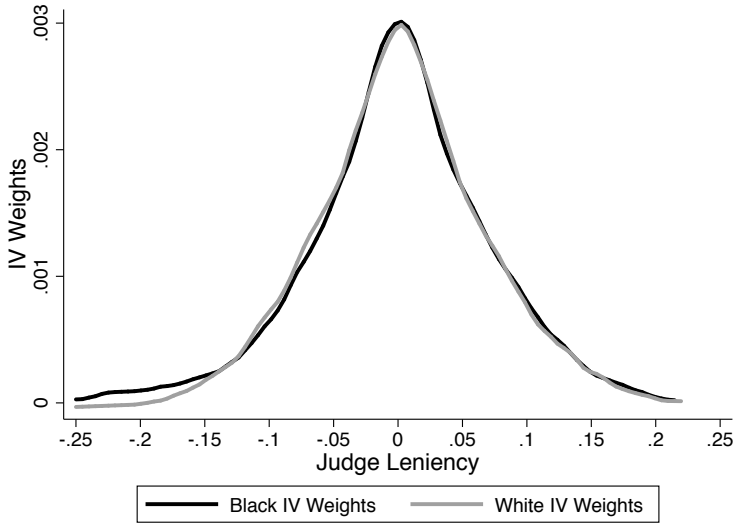
Intuitively, the MTE approach relies on identifying the MTE curve. To do so, we must impose structure on the relationship between the propensity score and the outcome of interest. This implies we also impose structure on the derivative of this relationship, which is equal to the MTE curve. If the structure does not bias our estimate of the MTE curve, then we should be able to construct the non-parametric IV by taking the weighted average of the MTE curve shown in Equation (63). However, if the estimated MTE is biased, then in general, the weighted average of the MTE will not be equal to the non-parametric IV estimate. We find that our MTE weighted by the IV weights is very close to the non-parametric IV estimate of racial bias. Specifically, the white IV estimate for the effect of release on rearrest is equal to 0.236, while the MTE weighted by the white IV weights yields an estimate of 0.261. Similarly, the black IV estimate for the effect of release on rearrest is equal to 0.014, while the MTE weighted by the black IV weights yields an estimate of 0.021. These results indicate that our MTE is likely to be correctly specified.

Online Appendix Table B1: Correlation between IV Weights and Observables

	White IV Weights x 100	Black IV Weights x 100
	(1)	(2)
Discrimination	0.424***	0.518***
	(0.066)	(0.062)
Philadelphia	0.117***	0.104***
	(0.016)	(0.016)
Case Load (100s)	0.004***	0.006***
	(0.001)	(0.001)
Average Leniency	0.044	0.000
	(0.055)	(0.054)
Experience	−0.000	0.002*
	(0.001)	(0.001)
Minority Judge	0.003	0.004
	(0.008)	(0.008)
Observations	552	552

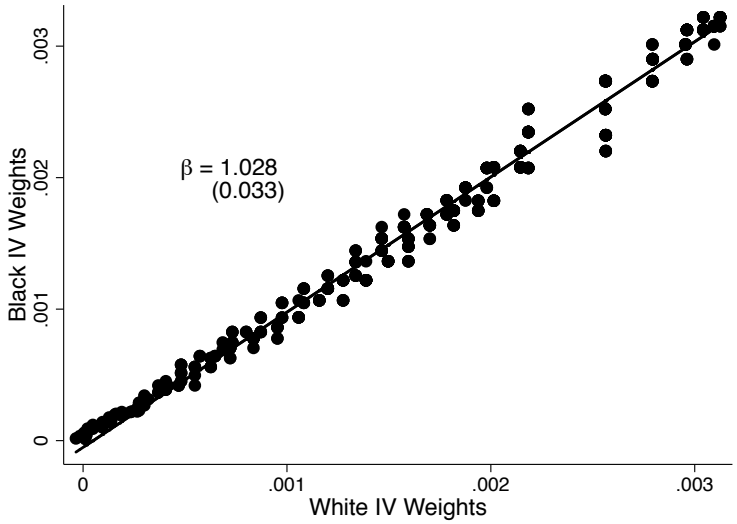
Notes. This table estimates the relationship between instrumental variable weights assigned to a given judge-by-year cell on observables of the judge-by-year cell. To ease readability, the coefficients are multiplied by a 100. Column 1 presents results for IV weights calculated for white defendants. Column 2 presents results for IV weights calculated for black defendants. To compute the weight assigned to a judge-by-year cell, we first compute the continuous weights by constructing sample analogues to the terms which appear in Equation (24) following the procedure described in Cornelissen et al. (2016) and Appendix B. To move from the continuous weights to a weight associated with a given judge, we compute the average leniency of each judge-by-year cell in the data. We then compute the weight associated with the average leniency of the judge-by-year cell using the results from the continuous weights estimation. We divide the resulting weights by the sum total to ensure the discretized weights sum to one.

Online Appendix Figure B1: Distribution of IV Weights by Race Across Judge Leniency Distribution



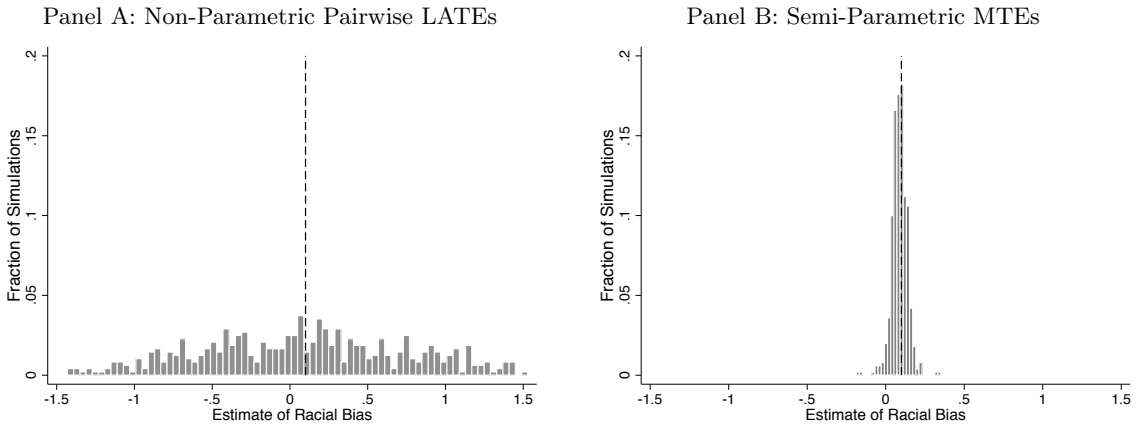
Note: This figure plots the instrumental variables weights over the distribution of judge leniency for both black and white defendants. To compute the instrumental variable weights, we first compute the continuous weights by constructing sample analogues to the terms which appear in Equation (24) following the procedure described in Cornelissen et al. (2016) and Online Appendix B. To move from the continuous weights to a weight associated with a given judge-by-year, we compute the average leniency of each judge-by-year cell in the data. We then compute the weight associated with the average leniency of the judge-by-year cell using the results from the continuous weights estimation. We divide the resulting weights by the sum total to ensure the discretized weights sum to one.

Online Appendix Figure B2: Correlation Between White IV Weights and Black IV Weights



Note: This figure plots the instrumental variables weight assigned to judge j in year t in the white leniency distribution vs. the instrumental variables weight assigned to judge j in year t in the black distribution. To compute the weight assigned to a judge-by-year cell, we first compute the continuous weights by constructing sample analogues to the terms which appear in Equation (24) following the procedure described in Cornelissen et al. (2016) and Online Appendix B. To move from the continuous weights to a weight associated with a given judge, we compute the average leniency of each judge-by-year cell in the data. We then compute the weight associated with the average leniency of the judge-by-year cell using the results from the continuous weights estimation. We divide the resulting weights by the sum total to ensure the discretized weights sum to one.

Online Appendix Figure B3: Monte Carlo Simulations of Racial Bias Estimators



Note: This figure reports the distribution of estimated racial bias using a race-specific judge leniency measure in simulated data with a "true" level of racial bias of 0.100. The simulated data include 170 judges, where each judge is assigned 500 black defendants and 500 white defendants. Defendant risk in the simulated data is drawn from a uniform distribution between 0 and 1. Judges release defendants if the risk is less than a judge-specific threshold, where the distribution of judge-specific threshold matches the empirical distribution of judge leniency. For each judge, we impose a 10 percentage point higher threshold for black defendants, so that the "true" level of racial bias in the simulated data is equal to 0.100. Panel A presents estimates from a non-parametric LATE procedure, where we form the Wald estimator between judge j and judge $j-1$ to estimate the release threshold for judge j. Panel B presents estimates from the MTE procedure. The estimate of racial bias is equal to the average estimated release threshold for white defendants minus the average estimated release threshold for black defendants across judges.

Online Appendix C: Simple Graphical Example

In this Online Appendix, we use a series of simple graphical examples to illustrate how a judge IV estimator for racial bias improves upon the standard OLS estimator. We first consider the OLS estimator in settings with either a single race-neutral judge or a single racially biased judge, showing that the standard estimator suffers from infra-marginality bias whenever there are differences in the risk distributions of black and white defendants. We then use a simple two-judge example to illustrate how a judge IV estimator can alleviate the infra-marginality bias in both settings.

C.1. OLS Estimator

To illustrate the potential for infra-marginality bias when using a standard OLS estimator, we begin with the case of a single race-neutral judge. The judge perfectly observes risk and chooses the same threshold for white and black defendants, but the distributions of risk differ by defendant race. Panel A of Online Appendix Figure C1 illustrates such a case, where we assume that white defendants have more mass in the left tail of the risk distribution, i.e. that whites are, on average, less risky than blacks. Letting the vertical lines denote the judge's release threshold, standard OLS estimates of α_W and α_B measure the average risk of released defendants for white and black defendants, respectively. In the case illustrated in Panel A, the standard OLS estimator indicates that the judge is biased against white defendants, when, in reality, the judge is race-neutral.

To further illustrate this point, Panel B of Online Appendix Figure C1 considers the case of a single judge that is racially biased against black defendants. Once again, the distributions of risk differ by defendant race, but now the judge chooses different thresholds for white and black defendants. In the case illustrated in Panel B, white and black defendants have the exact same expected risk conditional on release. As a result, the standard OLS estimator indicates that the judge is race-neutral, when, in reality, the judge is biased against black defendants. Following the same logic, we could choose risk distributions and release thresholds such that the OLS estimator indicates racial bias against white defendants or racial bias against black defendants. In other words, the OLS estimator is uninformative about the extent of racial bias in bail decisions without strong assumptions about differences in the underlying distributions of risk by defendant race.

C.2. IV Estimators

We now illustrate how a judge IV estimator for racial bias can potentially solve this infra-marginality problem by focusing the analysis on defendants at the margin of release. We use a simple two-judge example to illustrate the intuition behind our approach, while maintaining our assumption that judges perfectly observe risk and that the distributions of risk differ by defendant race. Throughout, we assume that judge 2 is more lenient than judge 1.

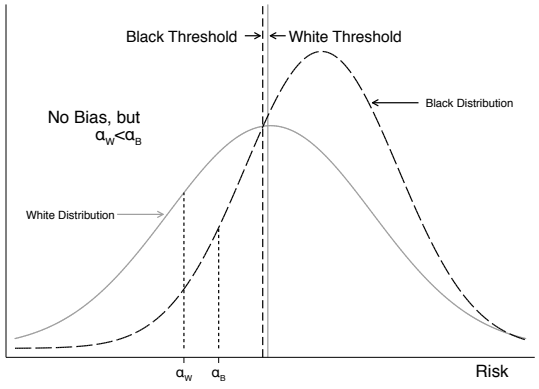
Panel C of Online Appendix Figure C1 considers the case where both judges are race-neutral, such that both judges use the same thresholds of release for white and black defendants. In this case, an IV estimator using judge leniency as an instrument for pre-trial release will estimate the average

risk for defendants who are released by the lenient judge but detained by the strict judge (i.e. the average risk of compliers), α_W^{IV} and α_B^{IV}. When the two judges are "close enough" in leniency, the IV estimator for racial bias will approximately estimate the risk of marginally released black defendants and marginally released white defendants. Intuitively, the IV estimator measures misconduct risk only for defendants near the margin of release, essentially ignoring the risk of defendants who are infra-marginal to the judge thresholds. As our measure of judge leniency becomes more continuous, our IV estimator will consistently estimate racial bias as the difference between α_W^{IV} and α_B^{IV}. The IV estimator will therefore indicate that marginal black and marginal white defendants have similar misconduct rates, allowing us to correctly conclude that the judges are race-neutral.

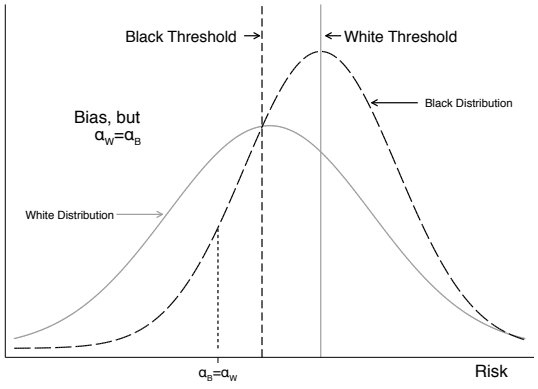
To further illustrate this point, Panel D of Online Appendix Figure C1 considers the case where both judges are racially biased against black defendants, such that both judges have higher thresholds of release for white defendants relative to black defendants. Following the same logic as above, the IV estimator measures the pre-trial misconduct risk of marginally released white and black defendants, α_W^{IV} and α_B^{IV}, so long as the two judges are "close enough" in leniency. The IV estimator will therefore indicate that marginal black defendants are lower risk than marginal white defendants, allowing us to correctly conclude that judges are racially biased against black defendants.

Online Appendix Figure C1: Infra-marginality Bias with OLS and Judge IV Estimators

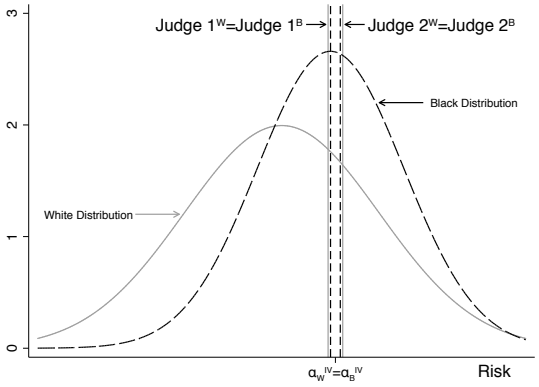
Panel A: OLS Estimator with Race-Neutral Judge



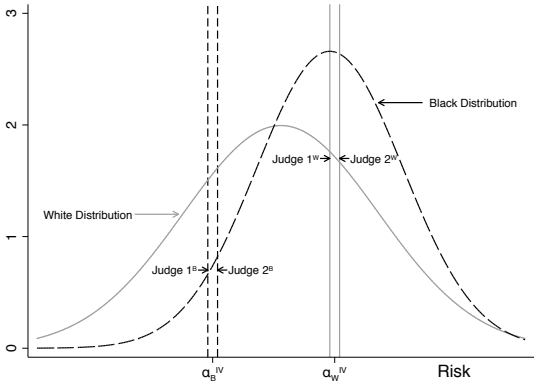
Panel B: OLS Estimator with Biased Judge



Panel C: IV Estimator with Two Race-Neutral Judges



Panel D: IV Estimator with Two Biased Judges



Note: These figures plot hypothetical risk distributions for white and black defendants. Panel A illustrates the OLS estimator with a race-neutral judge that chooses the same threshold for release for white and black defendants. Panel B illustrates the OLS estimator with a racially biased judge that chooses a higher threshold for release for white defendants compared to black defendants. Panel C illustrates the judge IV estimator with two race-neutral judges. Panel D illustrates the judge IV estimator with two racially biased judges.

Online Appendix D: Data Appendix

Judge Leniency: We calculate judge leniency as the leave-out mean residualized pre-trial release decisions of the assigned judge within a bail year. We use the residual pre-trial release decision after removing court-by-time fixed effects. In our main results, we define pre-trial release based on whether a defendant was ever released prior to case disposition.

Release on Recognizance: An indicator for whether the defendant was released on recognizance (ROR), where the defendant secures release on the promise to return to court for his next scheduled hearing. ROR is used for defendants who show minimal risk of flight, no history of failure to appear for court proceedings, and pose no apparent threat of harm to the public.

Non-Monetary Bail w/Conditions: An indicator for whether the defendant was released on non-monetary bail with conditions, also known as conditional release. Non-monetary conditions include monitoring, supervision, halfway houses, and treatments of various sorts, among other options.

Monetary Bail: An indicator for whether the defendant was assigned monetary bail. Under monetary bail, a defendant is generally required to post a bail payment to secure release, typically 10 percent of the bail amount, which can be posted directly by the defendant or by sureties such as bail bondsmen.

Bail Amount: Assigned monetary bail amount in thousands, set equal to zero for defendants who receive non-monetary bail with conditions or ROR.

Race: Indicator for whether the defendant is black (versus non-black).

Hispanic: We match the surnames in our data to census genealogical records of surnames. If the probability a given surname is Hispanic is greater than 70 percent, we label the defendant as Hispanic.

Prior Offense in Past Year: An indicator for whether the defendant had been charged for a prior offense in the past year of the bail hearing within the same county, set to missing for defendants who we cannot observe for a full year prior to their bail hearing.

Arrested on Bail in Past Year: An indicator for whether the defendant had been arrested while out on bail in the past year within the same county, set to missing for defendants who we cannot observe for a full year prior to their bail hearing.

Failed to Appear in Court in Past Year: An indicator for whether the defendant failed to appear in court while out on bail in the past year within the same county, set to missing for defendants who we cannot observe for a full year prior to their bail hearing. This variable is only available in Philadelphia.

Number of Offenses: Total number of charged offenses.

Felony Offense: An indicator for whether the defendant is charged with a felony offense.

Misdemeanor Offense: An indicator for whether the defendant is charged with only misdemeanor offenses.

Any Drug Offense: An indicator for whether the defendant is charged with a drug offense.

Any DUI Offense: An indicator for whether the defendant is charged with a DUI offense.

Any Violent Offense: An indicator for whether the defendant is charged with a violent offense.

Any Property Offense: An indicator for whether the defendant is charged with a property offense.

Rearrest Prior to Disposition: An indicator for whether the defendant was rearrested for a new crime prior to case disposition.

Failure to Appear in Court: An indicator for whether the defendant failed to appear for a required court appearance, as proxied by the issuance of a bench warrant. This outcome is only available in Philadelphia.

Failure to Appear in Court or Rearrest Prior to Disposition: An indicator for whether a defendant failed to appear in court or was rearrested in Philadelphia, and for whether a defendant was rearrested in Miami.

Judge Race: We collect information on judge race from court directories and conversations with court officials. All judges in Philadelphia are white. Information on judge race in Miami is missing for two of the 170 judges in our sample.

Judge Experience: We use historical court records back to 1999 to compute experience, which we define as the difference between bail year and start year (earliest 1999). In our sample, years of experience range from zero to 15 years.

Online Appendix E: Institutional Details

The institutional details described in this Online Appendix follow directly from Dobbie et al. (2018). Like the federal government, both Pennsylvania and Florida grant a constitutional right to some form of bail for most defendants. For instance, Article I, §14 of the Pennsylvania Constitution states that "[a]ll prisoners shall be bailable by sufficient sureties, unless for capital offenses or for offenses for which the maximum sentence is life imprisonment or unless no condition or combination of conditions other than imprisonment will reasonably assure the safety of any person and the community...." Article I, §14 of the Florida Constitution states that "[u]nless charged with a capital offense or an offense punishable by life imprisonment...every person charged with a crime...shall be entitled to pretrial release on reasonable conditions."

Philadelphia County: In Philadelphia County, defendants are brought to one of six police stations immediately following their arrest, where they are interviewed by the city's Pre-Trial Services Bail Unit. The Philadelphia Bail Unit interviews all adults charged with offenses in Philadelphia through videoconference, collecting information on each defendant's charge severity, personal and financial history, family or community ties, and criminal history. The Bail Unit then uses this information to generate a release recommendation based on a four-by-ten grid of bail guidelines that is presented to the bail judge at the bail hearing. However, these bail guidelines are only followed by the bail judge about half the time, with judges often imposing monetary bail instead of the recommended non-monetary options (Shubik-Richards and Stemen 2010).

After the Pre-Trial Services interview is completed and the charges are approved by the Philadelphia District Attorney's Office, defendants are brought in for a bail hearing. Bail hearings are conducted through videoconference by the bail judge on duty, with representatives from both the district attorney and local public defender's offices (or private defense counsel) present. However, while a defense attorney is present at the bail hearing, there is usually no real opportunity for defendants to speak with the attorney prior to the hearing. At the hearing itself, the bail judge reads the charges against the defendant, informs the defendant of his right to counsel, sets bail after hearing from representatives from the prosecutor's office and the defendant's counsel, and schedules the next court date. After the bail hearing, the defendant has an opportunity to post bail, secure counsel, and notify others of the arrest. If the defendant is unable to post bail, he is detained but has the opportunity to petition for a bail modification in subsequent court proceedings.

Under the Pennsylvania Rules of Criminal Procedure, "the bail authority shall consider all available information as that information is relevant to the defendant's appearance or nonappearance at subsequent proceedings, or compliance or noncompliance with the conditions of the bail bond," including information such as the nature of the offense, the defendant's employment status and relationships, and whether the defendant has a record of bail violations or flight. Pa. R. Crim. P. 523. In setting monetary bail, "[t]he amount of the monetary condition shall not be greater than is necessary to reasonably ensure the defendant's appearance and compliance with the conditions of the bail bond." Pa. R. Crim. P. 524. Under Pa. R. Crim. 526, a required condition of any

bail bond is that the defendant "refrain from criminal activity." In Philadelphia, it is well known that bail judges consider the risk of new crime when setting bail (see Goldkamp and Gottfredson 1988), and in fact, the Philadelphia bail guidelines are designed to "reduce the risk of releasing dangerous defendants into the community while ensuring that defendants who pose minimal risk are not confined to prison to await trial."[3]

Miami-Dade County: The Miami-Dade bail system follows a similar procedure, with one important exception. As opposed to Philadelphia where all defendants are required to have a bail hearing, most defendants in Miami-Dade can be immediately released following arrest and booking by posting an amount designated by a standard bail schedule. The standard bail schedule ranks offenses according to their seriousness and assigns an amount of bond that must be posted before release. Critics have argued that this kind of standardized bail schedule discriminates against poor defendants by setting a fixed price for release according to the charged offense rather than taking into account a defendant's ability to pay, or propensity to flee or commit a new crime. Approximately 30 percent of all defendants in Miami-Dade are released prior to a bail hearing through the standard bail schedule, with the other 70 percent of defendants attending a bail hearing (Goldkamp and Gottfredson 1988).

If a defendant is unable to post the standard bail amount in Miami-Dade, there is a bail hearing within 24 hours of arrest where defendants can argue for a reduced bail amount. Miami-Dade conducts separate daily hearings for felony and misdemeanor cases through videoconference by the bail judge on duty. At the bail hearing, the court will determine whether or not there is sufficient probable cause to detain the arrestee and if so, the appropriate bail conditions. The standard bail amount may be lowered, raised, or remain the same as the standard bail amount depending on the case situation and the arguments made by defense counsel and the prosecutor. While monetary bail amounts at this stage often follow the standard bail schedule, the choice between monetary versus non-monetary bail conditions varies widely across judges in Miami-Dade (Goldkamp and Gottfredson 1988).

Under the Florida Rules of Criminal Procedure, "[t]he judicial officer shall impose the first ... conditions of release that will reasonably protect the community from risk of physical harm to persons, assure the presence of the accused at trial, or assure the integrity of the judicial process." Fl. R. Crim. P. 3.131. As noted in Florida's bail statute, "[i]t is the intent of the Legislature that the primary consideration be the protection of the community from risk of physical harm to persons." Fla. Stat. Ann. §907.041(1).

Institutional Features Relevant to the Empirical Design: Our empirical strategy exploits variation in the pre-trial release tendencies of the assigned bail judge. There are three features of the Philadelphia and Miami-Dade bail systems that make them an appropriate setting for our research design. First, there are multiple bail judges serving simultaneously, allowing us to measure variation in bail decisions across judges. At any point in time, Philadelphia has six bail judges that only make bail decisions. In Miami-Dade, weekday cases are handled by a single bail judge, but weekend cases are

[3]See https://www.courts.phila.gov/pdf/notices/2012/6-12-12-Notice-to-Bar-Proposed-Bail-Guidelines.pdf.

handled by approximately 60 different judges on a rotating basis. These weekend bail judges are trial court judges from the misdemeanor and felony courts in Miami-Dade that assist the bail court with weekend cases.

Second, the assignment of judges is based on rotation systems, providing quasi-random variation in which bail judge a defendant is assigned to. In Philadelphia, the six bail judges serve rotating eight-hour shifts in order to balance caseloads. Three judges serve together every five days, with one bail judge serving the morning shift (7:30AM–3:30PM), another serving the afternoon shift (3:30PM–11:30PM), and the final judge serving the night shift (11:30PM–7:30AM). In Miami-Dade, the weekend bail judges rotate through the felony and misdemeanor bail hearings each weekend to ensure balanced caseloads during the year. Every Saturday and Sunday beginning at 9:00AM, one judge works the misdemeanor shift and another judge works the felony shift.

Third, there is very limited scope for influencing which bail judge will hear the case, as most individuals are brought for a bail hearing shortly following the arrest. In Philadelphia, all adults arrested and charged with a felony or misdemeanor appear before a bail judge for a formal bail hearing, which is usually scheduled within 24 hours of arrest. A defendant is automatically assigned to the bail judge on duty. There is also limited room for influencing which bail judge will hear the case in Miami-Dade, as arrested felony and misdemeanor defendants are brought in for their hearing within 24 hours following arrest to the bail judge on duty.

Online Appendix F: Model of Stereotypes

In this Online Appendix, we consider whether a model of stereotypes can generate the pre-trial release rates we observe in our data. To do so, we assume a functional form for how judges form perceptions of risk and ask if this model can match the patterns we observe in the data.

F.1. Calculating Predicted Risk:

We begin by estimating predicted risk using a machine learning algorithm that efficiently uses all observable crime and defendant characteristics. In short, we use a randomly-selected subset of the data to train the model using all individuals released on bail. In training the model, we must choose the shrinkage, the number of trees, and the depth of each tree. Following common practice, we choose the smallest shrinkage parameter (i.e. 0.005) that allows the training process to run in a reasonable time frame. We use a 5-fold cross validation on the training sample in order to choose the optimal number of trees for the predictions. The interaction depth is set to 5, which allows each tree to use at most 5 variables. Using the optimal number of trees from the cross validation step, predicted probabilities are then created for the full sample.

Following the construction of the continuous predicted risk variable, we split the predicted risk measure into 100 equal sized bins. One potential concern with this procedure is that observably high-risk defendants may actually be low-risk based on variables observed by the judges, but not by the econometrician. To better understand the importance of this issue, we follow Kleinberg et al. (2018) and plot the relationship between predicted risk and true risk in the test sample. We find that predicted risk is a strong predictor of true risk, indicating that the defendants released by judges do not have unusual unobservables which make their outcomes systematically diverge from what is expected (see Online Appendix Figure A3). This is true for both white and black defendants. Therefore, we interpret the predicted distributions of risk based on observables as the true distributions of risk throughout.

F.2. No Stereotypes Benchmark:

Following the construction of our predicted risk measure, we compute the fraction of black defendants that would be released if they were treated the same as white defendants. This calculation will serve as a benchmark for the stereotype model discussed below. To make this benchmark calculation, we assume judges accurately predict the risk of white defendants so that we can generate a relationship between release and risk, which we can then apply to black defendants. Under this assumption, we find that the implied release rate for black defendants is 70.8 percent if they were treated the same as white defendants. This implied release rate is lower than the true release rate of white defendants (71.1 percent), but higher than the true release rate for black defendants (68.8 percent), consistent with our main finding that judges over-detain black defendants.

F.3. Model with Stereotypes:

We can now consider whether a simple model of stereotypes can rationalize the difference in true release rates. Following Bordalo et al. (2016), we assume judges form beliefs about the distribution of risk through a representativeness-based discounting model. Basically, the weight attached to a given risk type t is increasing in the representativeness of t. Formally, let $\pi_{t,r}$ be the probability that a defendant of race r is in risk category $t \in \{1, ..., 100\}$. In our data, a defendant with $t = 1$ has a 2.7 percent expected probability of being rearrested before disposition while a defendant with $t = 100$ has a 74.5 percent probability of being rearrested before disposition.

Let $\pi^{st}_{t,r}$ be the stereotyped belief that a defendant of race r is in risk category t. The stereotyped beliefs for black defendants, $\pi^{st}_{t,B}$, is given by:

$$\pi^{st}_{t,B} = \pi_{t,B} \frac{\left(\frac{\pi_{t,B}}{\pi_{t,W}}\right)^{\theta}}{\sum_{s \in T} \pi_{s,B} \left(\frac{\pi_{s,B}}{\pi_{s,W}}\right)^{\theta}} \tag{64}$$

where θ captures the extent to which representativeness distorts beliefs and the representativeness ratio, $\frac{\pi_{t,B}}{\pi_{t,W}}$, is equal to the probability a defendant is black given risk category t divided by the probability a defendant is white given risk category t. Recall from Figure III that representativeness of blacks is strictly increasing in risk. Therefore, a representativeness-based discounting model will over-weight the right tail of risk for black defendants.

To compute the stereotyped distribution, we first assume a value of θ, and then compute $\pi_{t,r}$ for every risk category t and race r. We can then compute $\pi^{st}_{t,B}$ by plugging in the values for $\pi_{t,r}$ and the assumed value of θ into Equation (64).

From the distribution of $\pi^{st}_{t,B}$, we compute the implied average release rate by multiplying the fraction of defendants believed to be at a given risk level by the probability of release for that risk level and summing up over all risk levels. Formally,

$$\mathbb{E}[Released_i = 1 | r_i = B] = \sum_{s=1}^{100} \pi^{st}_{s,B} \mathbb{E}[Released_i = 1 | t = s, r_i = B] \tag{65}$$

In the equation above, we cannot compute $\mathbb{E}[Released_i = 1 | t = s, r_i = B]$ given that we explicitly assume judges make prediction errors for black defendants. That is, we do not know at what rate judges would release black defendants with risk equal to s, given that judges do not accurately predict risk for black defendants. However, in a stereotypes model, we can replace $\mathbb{E}[Released_i = 1 | t = s, r_i = B] = \mathbb{E}[Released_i = 1 | t = s, r_i = W]$ (i.e. given that if there is no taste-based discrimination, then conditional on perceived risk, the release rate will be equal between races). Under our additional assumption that judges accurately predict the risk of whites, we can estimate $\mathbb{E}[Released_i = 1 | t = s, r_i = W]$ for all s. Therefore, we can compute every value on the right hand side of Equation (65), from which we can back out the average release rate for black defendants from the stereotyped distribution.

We find that $\theta = 1.9$ rationalizes the average release rate for blacks we observe in the data (68.8 percent). That is, if judges use a representativeness-based discounting model with $\theta = 1.9$ to form perceptions of the risk distribution, we would expect judges to release 68.8 percent of all black defendants. To understand how far these beliefs are from the true distribution of risk, we plot the stereotyped distribution for blacks with $\theta = 1.9$ alongside the true distribution of risk for blacks in Online Appendix Figure A4. The average risk in the stereotyped distribution is about 5.4 percentage points greater than the mean in the true distribution of risk.