Type I and Type II Error Probabilities in the Courtroom

Shin Kanaya[1] and Luke Taylor[2]

Abstract

We estimate the likelihood of miscarriages of justice by reframing the problem in the context of misclassified binary choice models. Our estimator is based on new nonparametric identification results, for which we provide methods to empirically test the key identifying assumptions and alternative identification schemes for when these checks fail. Using case-level data from Virginia, we find blacks have both a higher probability of conviction when innocent and a higher probability of acquittal when guilty, relative to whites. We go on to show that this seemingly contradictory result is, in fact, consistent with a model where blacks are discriminated against at both the arrest and the conviction stage of the judicial process.

JEL Classification Codes: K14; K41; C14; C25.

[1]Department of Economics, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ, United Kingdom. Email: shin.kanaya@essex.ac.uk.
[2]Department of Economics, Aarhus University, Fuglesangs Alle 4, Aarhus V 8210, Denmark. Email: lntaylor@econ.au.dk.

1 Introduction

Criminal courts have existed for more than a thousand years, yet still there exists no reliable method to measure their performance. This is not due to apathy. Miscarriages of justice have been a subject of debate going back to antiquity (Zalman, 2017) and, in recent years, this topic has seen a surge of interest. Indeed, the popularity of television shows and podcasts chronicling the (in)famous trials of O.J. Simpson, Steven Avery, and the 'Central Park Five' among others, have thrust the fallibility of courts further into the spotlight.

Judicial errors are not a matter of mere curiosity though. Since its inception in 1992, the Innocence Project has aided in the exoneration of 367 innocent people, who had collectively served more than 5000 years in prison.[1] Moreover, 21 of those individuals were sentenced to death. This begs the question: how many people have been wrongfully executed, paying the ultimate price for a type I error? The financial cost borne by the state through unnecessarily incarcerating innocent inmates is also not small; a single prisoner costs an average of $31 000 per year to keep behind bars.[2] For type II errors, i.e. failing to punish a guilty defendant, the costs fall predominantly on the victims who do not find the justice they deserve, as well as on society more generally through releasing criminals back into the population and through an erosion of belief in the justice system.

In this paper, we estimate these error rates for judges using data on more than five million court cases from Virginia. Our method is based on reframing the problem in the context of misclassified binary choice models where the misclassification rates can be interpreted as type I and type II errors, respectively. In particular, we consider the decision of the judge - to convict or acquit - as a noisy measure of the true guilt of the defendant. We give new nonparametric identification results for the misclassified binary choice model which admit simple and intuitive estimators. These estimators are of independent interest and have potential applications in various contexts, as discussed later in this section. We also provide methods to empirically test the identifying assumptions and give alternative estimation schemes for cases which fail these tests.

The key identification condition requires a continuous regressor which affects the true outcome but which does not affect the misclassification rates. In our justice context, this conditional mean independence assumption translates to a regressor which affects the probability that the defendant is guilty but does not affect the probability of a miscarriage of justice. The regressor must also satisfy a large-support condition similar to that of Lewbel (2000b). A variable of this type is known as a

[1] The Innocence Project uses DNA evidence to overturn wrongful convictions.
[2] Vera Institute of Justice.

'special regressor' (Lewbel, 1998).

As our special regressor, we use a measure of future criminality (constructed using gradient boosted regression trees). First, empirical evidence that this variable satisfies the large-support condition is provided. We then go on to test the conditional mean independence assumption. Intuitively, if a judge were informed, for example, that a defendant will be convicted for the same type of crime many times in the future, this would provide substantial information regarding the defendant's likelihood of guilt for the current crime. However, since future criminality does not materialise until after the trial, there is no means for it to affect the court proceedings.

Nonetheless, there are still two concerns. First, conditional on true guilt and a set of control variables, unobservable characteristics which are related to future criminality must not be related to the misclassification rates. Second, the court's ruling must not affect future criminality.

For the former, by controlling for previous criminality, among other variables, it is hoped that many potential sources of correlation between the misclassification rates and future criminality are captured. Furthermore, we argue that if the unobservable variable affects the probability of a miscarriage of justice through a channel other than true guilt, this variable must represent a bias in the judge's decision. Consequently, it is unlikely to be related to the future criminal behaviour of the defendant.

For the latter concern, to mute the effect of conviction on future criminality, the sample is restricted to non-felony crimes. Although this removes some of the more interesting offences, it is possible that the less serious cases are likely to be subject to more frequent miscarriages of justice: the conveyor-belt of defendants passing through the justice system results in judges and lawyers giving less time to each case.

Having restricted the sample, we test the plausibility of the court decision having no effect on future criminality. To this end, we use the random assignment of judges to cases as an instrumental variable to show that the effect of conviction on future criminality is small and statistically insignificant. Together, these results suggest that the primary identification conditions hold in our setting.

We find that the probability of convicting an innocent defendant is relatively high, ranging from 16% to 28%, depending on race and gender. However, this should not be confused with the probability of convicting an innocent *person*. The dataset used contains only individuals who have been arrested for a given crime. Thus, the results are conditional on having been arrested, which greatly increases the probability of conviction relative to a member of the general public. For the probability of acquitting a guilty defendant, the estimates range from 16% to 19%, depending on race and gender.

We also find that males face a higher probability of being convicted when innocent and a lower

probability of being acquitted when guilty, relative to females. This suggests that the strength of evidence required to convict a man is lower than that for a woman.

Interestingly, in comparison to whites, blacks are more likely to be convicted for a crime they did not commit but are also more likely to be acquitted for a crime they did commit. To answer this seemingly contradictory result, we develop a theoretical model of the decision process of both the police and judge. By calibrating the parameters of this simple model using the empirical estimates obtained, we show that both the arrest and conviction thresholds are generally lower for blacks relative to whites. While this does not prove the existence of racial animus, it shows that our empirical results are largely consistent with model-based predictions of discrimination against blacks.

Although our focus is on errors in the courtroom, our estimation strategy is applicable more generally. Whenever a choice is made with incomplete information, a type I or type II error is possible. However, as with miscarriages of justice, in many cases, it is never known if a mistake has been made. For example, interviews for prospective employees are designed to gather information about the applicant to avoid either hiring an inadequate worker or passing on a suitable one. In some circumstances, it is possible to discover if the wrong person has been hired. However, since no new information is gathered after the decision, it is almost always impossible to know if the right person was not hired. This is also true of university admissions, promotions, and lending applications, to name only a few examples.

In general, our approach is useful for situations where the type I or type II error is either impossible or expensive to observe directly. As an example of the latter, consider the case of fraudulent insurance and welfare benefits claims. In this context, it is possible to determine if a claim is dishonest by a careful examination of the case. However, this can be expensive and time-consuming. Although the techniques put forward in this paper cannot determine whether an *individual* claim is fraudulent, they can be used to predict which claims have a higher probability of being false based on observable characteristics. This would allow investigators to target their efforts to detect fraud more intelligently.

1.1 Previous Literature

This paper contributes to two main literatures. First, we add to work on estimating the prevalence of miscarriages of justice. Almost without exception, the existing research in this area has been restricted to estimating the probability of wrongful conviction - as opposed to wrongful acquittal. Furthermore, this work has, almost exclusively, used data on exonerations (see, for example, Risinger, 2006; Gross and O'Brien, 2008; Gross, O'Brien, Hu and Kennedy, 2014). However, an exoneration is not equivalent

to innocence.

First, the number of exonerations is likely to represent only a small fraction of the total number of false convictions. In many cases, the effort to uncover these miscarriages of justice is not made; the severity of the crime, and hence the punishment, is too low to warrant the use of limited resources. Moreover, even if an investigation is conducted, it may be that the evidence required to overturn a previous conviction does not exist.

Second, in the majority of cases, exonerations occur due to misconduct during the arrest or trial.[3] Former Supreme Justice Antonin Scalia explains, "most [overturned death sentences] are based on legal errors that have little or nothing to do with guilt".[4]

These limitations of exoneration data are well documented (see, for example, Acker, 2017) and have resulted in several approaches attempting to mitigate these shortcomings. A notable example is Gross *et al.* (2014) who restrict their analysis of exonerations to death row inmates only. It is reasoned that by considering a subset of convictions for which the majority of mistakes are identified, the estimate obtained will be more accurate than using the entire population of convictions. They find the probability of convicting an innocent defendant to be 4.1% but acknowledge that this is likely to be a lower bound for the actual probability. There are also reservations about extrapolating this to other cases; the judge may spend more time deliberating the evidence or be more likely to err on the side of caution when a person's life is at stake.

To the best of our knowledge, Spencer (2007) is the only other work - together with the present paper - to estimate the probability of acquitting a guilty defendant and which does not rely on exoneration data. His method is similar in style to ours and can be viewed in a measurement error framework. By analysing the rate of agreement between a judge and a jury, he shows that both the probability of a wrongful conviction and a wrongful acquittal can be estimated. However, he is transparent regarding the strong assumptions imposed for identification. In particular, judges and juries are assumed to make mistakes at the same rate. Furthermore, the probability of a correct decision from the judge is independent of the likelihood of a correct decision from the jury in a given trial. This seems unrealistic. For example, in a complex case, the probability of a correct decision is likely to be lower for both judge and jury. In contrast, we only require data on the decision of either a judge or jury - not both.[5]

Finally, Bjerk and Helland (2019) take a different approach and concede that while the exact

[3]National Registry of Exonerations.

[4]Kansas v Marsh, 548 US 163, 182 (2006).

[5]Our sample contains only cases heard by judges; however, the approach used in this paper is equally applicable to jury trials.

probability of a false conviction may be beyond reach, differences in the exoneration rate can shed light on racial discrepancies in sentencing. They find that the exoneration rate of white defendants for rape cases was less than two-thirds of that for corresponding black defendants.

The present paper also adds to the literature on misclassified binary choice models. The first identification results for this model were given by Hausman, Abrevaya and Scott-Morton (1998); however, their approach was restricted to a parametric model. Lewbel (2000b) extended their results to a semiparametric model and used a special regressor approach to estimate the parameters. We also use a special regressor approach, but in contrast to Lewbel (2000b), we provide weaker conditions under which the misclassification rates can be identified and propose a simpler estimator. Indeed, in that paper, he explains that ''the estimators provided here are not likely to be very practical, because they involve up to third-order derivatives and repeated applications of nonparametric regression'' (pp. 607-608). In contrast, our estimator uses only a single nonparametric regression and does not require the estimation of derivatives. Examples of other papers which apply special regressor methods include Heckman and Navarro (2007) in a dynamic choice model, Berry and Haile (2014) to estimate demand functions, and Lewbel and Tang (2015) and Khan and Nekipelov (2018) in game-theoretic models.

2 Identification

2.1 Baseline Identification

This section provides details of the general model, the identification strategies, and the required assumptions. The objects of interest are the type I and type II error probabilities defined as

$$\alpha_1(x) := P\left[Y = 1 | Y^* = 0, X = x\right] \text{ and } \alpha_2(x) := P\left[Y = 0 | Y^* = 1, X = x\right],$$

respectively, where Y and Y^* are binary-valued variables. Y^* denotes the true unobservable outcome, Y is an observed but misclassified version of Y^*, and X represents a vector of observable covariates. Finally, let $supp\left(Z\right)$ denote the support of a random vector Z.

We also suppose the availability of an additional (scalar) variable, V, which we describe as a special regressor. This regressor is assumed to satisfy some conditions that are distinct from the set of other covariates X. In particular, we assume the following:

Assumption 1 [Exclusion Restriction] There exists a scalar-valued, continuously distributed variable V which satisfies

$$E[Y|Y^*, X, V] = E[Y|Y^*, X] \text{ almost surely.}$$

Assumption 2 [Single-Index Structure] The true outcome Y^* and regressors (X, V) are related through

$$Y^* = \mathcal{I}(V + h(X) - U \geq 0), \tag{1}$$

where $\mathcal{I}(\cdot)$ is the indicator function, $h(\cdot)$ is an unknown scalar-valued function on the support of X, U is an unobservable random variable with $U \perp V|X$, and $U|X = x$ is continuously distributed for each x, i.e. the conditional cumulative distribution function (CDF) of U, $F_{U|X}(u|x)$, has a corresponding density $f_{U|X}(u|x)$.

In our empirical application, Y^* denotes whether the defendant is factually guilty (= 1) or innocent (= 0), Y indicates whether the defendant was convicted (= 1) or acquitted (= 0), and X includes information on both the case and the defendant. Hence, $\alpha_1(x)$ gives the probability of convicting an innocent defendant with characteristics x, and $\alpha_2(x)$ is the corresponding probability of acquitting a guilty defendant. Finally, future criminality of the defendant (defined in detail in Section 3.1) is used as the special regressor, V.

In this setup, an error made by the court is given by $(Y - Y^*)$ and, under Assumption 1, its conditional expectation can be written as

$$E[Y - Y^*|Y^*, X, V] = E[Y - Y^*|Y^*, X].$$

This says that, on average, the error depends on whether the defendant is factually guilty and on the characteristics of the case and the defendant, respectively. However, future criminality does not affect the error. In Section 4.1, we provide a thorough investigation of the validity of Assumption 1, including an instrumental variable analysis using the leniency of quasi-randomly assigned judges to cases as an instrument for conviction status.

The conditional independence condition $U \perp V|X$ in Assumption 2 resembles that of Lewbel (2000b, Assumption A2). This, together with the single index structure of Y^* in equation (1), leads

to the following expression for the 'conditional predictive probability' (CPP):

$$P\left[Y^* = 1 | \left(X, V\right) = (x, v)\right] = F_{U|X}\left(v + h(x) | x\right). \tag{2}$$

While Assumption 2 may look restrictive, it does not impose any significant restriction on the functional form of the CPP except for monotonicity in v. That is, any CPP that is monotone in v can be represented by the model in Assumption 2 under mild regularity conditions (cf. Theorem 3 of Magnac and Maurin, 2007, which gives a representation result for monotone binary choice models; see Appendix A for a detailed discussion). Note that this single-index model is not a structural model, i.e. it does not attempt to explain a defendant's criminal behaviour. It is merely a tool for the researcher to predict such behaviour retrospectively. This stands in contrast to previous work which uses similar single-index specifications and conditional independence assumptions to create structural models (see, for example, Berry and Haile, 2014), where careful consideration must be given to the underlying behavioural mechanisms which could result in such a model.

Together, Assumptions 1 and 2 allow us to write

$$\begin{aligned} P\left[Y = 1 | \left(X, V\right) = x, v\right] &= P\left[Y = 1 | Y^* = 0, \left(X, V\right) = x, v\right] P\left[Y^* = 0 | \left(X, V\right) = x, v\right] \\ &\quad + P\left[Y = 1 | Y^* = 1, \left(X, V\right) = x, v\right] P\left[Y^* = 1 | \left(X, V\right) = x, v\right] \\ &= \alpha_1(x)\left[1 - P\left[Y^* = 1 | \left(X, V\right) = x, v\right]\right] \\ &\quad + \left[1 - \alpha_2(x)\right] P\left[Y^* = 1 | \left(X, V\right) = x, v\right] \\ &= \alpha_1(x) + \left[1 - \alpha_1(x) - \alpha_2(x)\right] P\left[Y^* = 1 | \left(X, V\right) = x, v\right]. \end{aligned} \tag{3}$$

Note that the objects of interest, $\alpha_1(x)$ and $\alpha_2(x)$, are independent of v, while the left-hand side and the CPP depend on v. This expression forms the basis of our identification argument. However, one further assumption is required.

For clarity of exposition, we assume that the support of $V|X = x$ is given by a bounded and closed interval $[l_V^x, r_V^x]$ for each $x \in supp\left(X\right)$.[6]

[6]Note that all subsequent results carry over to other cases (with only slight modifications) which allow for an unbounded or (semi-) open interval, including $(-\infty, \infty)$.

Assumption 3 [Large Support Condition] For each $x \in supp\,(X)$,

$$\lim_{v \to l_V^x} P\left[Y^* = 1 | \,(X,V) = (x,v)\right] = 0, \tag{4}$$

$$\lim_{v \to r_V^x} P\left[Y^* = 1 | \,(X,V) = (x,v)\right] = 1. \tag{5}$$

Assumption 3 states that being factually guilty or not can be perfectly predicted by future criminality in its tail region. In other words, the support of V is sufficiently large: $[l_U^x, r_U^x] \subseteq [l_V^x, r_V^x]$, where $[l_U^x, r_U^x]$ is the support of $U|X = x$.[7]

From equation (3) and Assumption 3, we obtain

$$\lim_{v \to l_V^x} P\left[Y = 1 | \,(X,V) = (x,v)\right] = \alpha_1(x), \tag{6}$$

$$\lim_{v \to r_V^x} P\left[Y = 1 | \,(X,V) = (x,v)\right] = 1 - \alpha_2(x), \tag{7}$$

which establish the identification of $\alpha_1(x)$ and $\alpha_2(x)$, respectively. This type of identification is typically referred to as 'identification at infinity', particularly if the support of $V|X = x$ is $(-\infty, \infty)$ where $v \to l_V^x$ and $v \to r_V^x$ are replaced by $v \to -\infty$ and $v \to \infty$ in equations (6) and (7), respectively.

2.2 Testing the Large Support Condition

It is clear that the large support condition is critical to achieving identification. Thus, it is important to provide a method to empirically check the validity of this assumption.

The partial derivative of equation (3) with respect to v is given by

$$\frac{\partial}{\partial v} E\left[Y | V = v, X = x\right] = \left[1 - \alpha_1(x) - \alpha_2(x)\right] f_{U|X}(v + h(x)|x). \tag{8}$$

Note that the left-hand side of equation (8) is easily estimated from the data. Furthermore, for a given x, this partial derivative is a constant multiple of $f_{U|X}(v + h(x)|x)$. Thus, it is possible to evaluate this derivative in the interval $[l_V^x, r_V^x]$ to determine whether the tail condition is satisfied for a given x. If the derivative falls to zero in the upper tail, this suggests $F_{U|X}(r_V^x + h(x)|x) = 1$, providing that $f_{U|X}\,(\cdot|x)$ has no zero-probability intervals in the interior of its support. Equally, a zero derivative at the lower tail indicates $F_{U|X}(l_V^x + h(x)|x) = 0$. Consequently, the validity of the large support assumption can be easily checked for each tail condition and for each point of interest x.

[7] Assumption 3 is written using limit notation such that the conditions hold for unbounded or open settings. This reduces to $P\left[Y^* = 1 | \,(X,V) = (x, l_V^x)\right] = 0$ and $P\left[Y^* = 1 | \,(X,V) = (x, r_V^x)\right] = 1$ when the support of $V|X = x$ is $\left[l_V^x, r_V^x\right]$.

2.3 Relaxation of the Large Support Condition

Unfortunately, it may be difficult to satisfy the large support condition in many empirical settings. For example, in Section 4.2, the testing approach put forward in Section 2.2 is used to show that equation (5) of Assumption 3 is satisfied in our justice context, but equation (4) is not. Intuitively, there are defendants in the sample with high enough future criminality that it is possible to conclude they are guilty of the current crime. However, even for the most law-abiding citizens in the future, we cannot claim with any certainty that they are innocent of the current crime. Thus, only half of the large support assumption is satisfied.

As a result, it is worthwhile to pursue alternative identification mechanisms which do not require equations (4) and (5) to hold simultaneously. Without loss of generality, we proceed without the lower-bound condition (4) and impose the following assumptions:

Assumption 3′ [Alternative Large Support Condition] For each $x \in supp\,(X)$,

$$\lim_{v \to r_V^x} P\left[Y^* = 1 | \,(X, V) = (x, v)\right] \quad = \quad 1. \tag{9}$$

Assumption 4 [Mode-Median Coincidence/Limited Predictability] The conditional CDF $F_{U|X}\,(\cdot|x)$ is differentiable on the entire support of $U|X = x$ and has derivative $f_{U|X}\,(\cdot|x)$. There exists a unique maximum point (conditional mode) m_U^x of $f_{U|X}\,(\cdot|x)$ on $[l_V^x + h(x), r_U^x]$ which coincides with the conditional median of $U|X = x$, where r_U^x is the upper limit of the support of $U|X = x$.

Assumption 3′ is a weakening of Assumption 3 in that it removes the lower tail condition but maintains the upper (i.e. it is only supposed that $r_U^x \leq r_V^x$). Under Assumptions 1, 2, and 3′ it is possible to identify $\alpha_2(x)$ but not $\alpha_1(x)$.

However, Assumption 4 can be used to recover $\alpha_1(x)$. While a condition involving the conditional mode may look unusual for latent-variable discrete choice models, this assumption is satisfied by commonly-used parametric distributions. Its simplest sufficient condition is that $U|X = x$ is symmetric and single-peaked, as in the case of the Gaussian or logistic distributions; however, it does not exclude non-symmetric distributions.

It is also worthwhile to note that the maximum point m_U^x need not necessarily be the mode of $U|X = x$. That is, the true mode may exist outside of $[l_V^x + h(x), r_U^x]$. We simply require that the unique maximum point inside $[l_V^x + h(x), r_U^x]$ is equal to the median; nonetheless, we maintain the

mode interpretation for ease of understanding.

We interpret Assumption 4 as a limited predictability condition in the following way. From the form of the CPP in equation (2), the median value of $U|X = x$ occurs where the probability of the defendant being guilty is 0.5. Since we require $\text{Mode}\left[U|X = x\right] = \text{Median}\left[U|X = x\right]$, it must be that there is a significant proportion of defendants who are as likely to be guilty as they are to be innocent and, consequently, whose guilt is difficult for the researcher to predict.

It is also possible to interpret Assumption 4 as a type of location normalisation; Manski (1998) imposes a location normalisation through a conditional median restriction to identify $h(x) = x'\beta$. In the present context, his assumption corresponds to $\text{Median}\left[U|X\right] = 0$. While $\text{Mode}\left[U|X\right] = 0$ can play the same role, it is important to note that Manski (1998) considers an observable binary outcome. If Y^* were observable in our setting, either $\text{Mode}\left[U|X\right] = 0$ or $\text{Median}\left[U|X\right] = 0$ could be used to identify $h(x)$.[8,9] In this respect, Assumption 4 is stronger than necessary when Y^* is observable. Theorem 2 in Appendix A gives a representation result for the CPP and clarifies that indeed $\text{Mode}\left[U|X\right] = \text{Median}\left[U|X\right]$ imposes more structure on the CPP in equation (2) than either $\text{Mode}\left[U|X\right] = 0$ or $\text{Median}\left[U|X\right] = 0$. However, it appears that when Y^* is unobservable, some additional restriction, such as Assumption 4, must be imposed for identification.

We now illustrate how Assumption 4 can be used to restore the identification of $\alpha_1(x)$ when Assumption 3′ holds but Assumption 3 does not, i.e. when only the upper tail condition is satisfied. Recall that

$$P\left[Y = 1|\left(X, V\right) = (x, v)\right] = \alpha_1(x) + \left[1 - \alpha_1(x) - \alpha_2(x)\right] F_{U|X}\left(v + h(x)|x\right).$$

Taking the partial derivative with respect to v gives

$$\frac{\partial}{\partial v} P\left[Y = 1|\left(X, V\right) = (x, v)\right] = \left[1 - \alpha_1(x) - \alpha_2(x)\right] f_{U|X}\left(v + h(x)|x\right).$$

Since the right-hand side is a constant multiple of $f_{U|X}\left(v + h(x)|x\right)$ for a given x, if $\alpha_1(x) + \alpha_2(x) < 1$,

[8] It is not necessary to assume that these conditional measures are equal to 0. It is possible to use any known number $c_x \in \mathbb{R}$ for each x normalisation instead (see Theorem 2 in Appendix A).

[9] In this case, an additional condition would be required for identification: for the former mode condition, there must exist some v such that $P\left[Y^* = 1|(X, V) = (x, v)\right] = 1/2$ for each x; and for the latter median condition, $(\partial/\partial v)P\left[Y^* = 1|(X, V) = (x, v)\right]$ must have a unique maximiser v in the support of V for each x (see Theorem 2 in Appendix A).

we can write

$$\begin{aligned}\bar{v}(x) &:= \operatorname{argmax}_{v\in[l_V^x, r_V^x]} \frac{\partial}{\partial v} P\left[Y=1|\left(X,V\right)=(x,v)\right] \\ &= \operatorname{argmax}_{v\in[l_V^x, r_V^x]} f_{U|X}\left(v+h(x)|x\right). \end{aligned} \tag{10}$$

The restriction $\alpha_1(x)+\alpha_2(x)<1$ is what Hausman *et al.* (1998) call the monotonicity condition and is standard in the literature on misclassified binary variables. In our empirical setting, this states that the court's ruling is informative of the guilt of the defendant. If this did not hold, the court would make fewer mistakes if all those who were convicted were acquitted instead, and all those originally acquitted were now convicted; this seems an unlikely situation to occur.

Under Assumption 4, $[\bar{v}(x)+h(x)]$ is the median of $U|X=x$, so $F_{U|X}\left(\bar{v}(x)+h(x)|x\right)=1/2$. Therefore,

$$\begin{aligned} P\left[Y=1|\left(X,V\right)=(x,\bar{v}(x))\right] &= \alpha_1(x)+\left[1-\alpha_1(x)-\alpha_2(x)\right]/2 \\ &= \left[1+\alpha_1(x)-\alpha_2(x)\right]/2. \end{aligned}$$

This, together with equation (9) of Assumption 3′, allows for the identification of $\alpha_1(x)$ and $\alpha_2(x)$. We summarise this result in the following theorem.

Theorem 1 Suppose that Assumptions 1, 2, 3′, and 4 hold. If $\alpha_1(x)+\alpha_2(x)<1$ for each $x\in supp\left(X\right)$, then $\alpha_1(x)$ and $\alpha_2(x)$ are identified as

$$\begin{aligned}\alpha_1(x) &= 2P\left[Y=1|\left(X,V\right)=(x,\bar{v}(x))\right]-\lim_{v\to r_V^x} P\left[Y=1|\left(X,V\right)=(x,v)\right], \\ \alpha_2(x) &= 1-\lim_{v\to r_V^x} P\left[Y=1|\left(X,V\right)=(x,v)\right], \end{aligned}$$

where $\bar{v}(x)$ is defined in equation (10).

In principle, estimators for $\alpha_1(x)$ and $\alpha_2(x)$ can be constructed using empirical analogues of the expressions in Theorem 1. However, the following integral-based formula (derived in Appendix A) may be more practical:

$$\begin{aligned}\alpha_2(x) &= 1-P\left[Y=1|X=x\right] \\ &\quad -\int_{l_V^x}^{r_V^x} \left(\partial/\partial v\right) P\left[Y=1|\left(X,V\right)=(x,v)\right] F_{V|X}(v|x)dv. \end{aligned} \tag{11}$$

An estimator based on this object is likely to be more robust and allow an easier analysis of the asymptotic properties of the resulting estimator (cf. Goh, 2018).

3 Data

3.1 The Special Regressor

We open this section with a discussion of the choice of the special regressor: future criminal behaviour (defined in detail in Section 3.2). In Section 4.2, evidence is provided that this variable is highly correlated with the true guilt of the defendant and satisfies the large support condition. Crucially, we must also ensure that future criminality satisfies the conditional mean independence assumption. Since future criminality only becomes apparent after the judge's decision, there is no means for it to influence the outcome of the trial. Nonetheless, there are two concerns.

First, conditional on true guilt and a set of control variables, future criminality must be unrelated to unobservables which affect the probability of misclassification. Focussing on incorrect acquittal, recall that, essentially, our estimation strategy uses the release rate of individuals with very high future criminality as the false acquittal rate. To extrapolate this rate to individuals with lower levels of future criminality requires that all individuals share the same misclassification rate conditional on observable characteristics. We believe that by nonparametrically controlling for key variables such as race, gender, and previous criminality, we close off any potential channels through which unobserved characteristics could lead to different misclassification rates.

However, if there is still concern regarding a particular unobserved variable, the interpretation of the misclassification rate must be altered; the misclassification rate becomes specific to the group of individuals who have high future criminality. For example, if those with visible gang tattoos, which are not observed by the researcher, are more likely to have high future criminality and also have a different wrongful acquittal rate (conditional on control variables), then the misclassification rate becomes specific to those with visible gang tattoos.

A second issue is whether the judge's decision can affect the future criminality of the individual. If this is the case, future criminal behaviour and the court ruling will not be mean independent. For example, suppose that a conviction causes an individual to commit many crimes in the future. Our results would show that the court never convicts an innocent defendant because the convicted defendants have very high future criminality suggesting they are guilty of the current crime.

To mitigate concerns of this nature, the sample is first restricted to infractions and misdemeanours,

i.e. felonies are excluded, in order to reduce the potential effect of a conviction. In Virginia, infractions carry no prison sentence and a maximum fine of $250, while misdemeanours can be punished by a maximum prison sentence of one year. The mean prison sentence for those convicted in our sample is 14 days, and the median is no prison time at all. Nonetheless, we go on to formally test whether this assumption is satisfied in our data.

There is an abundance of previous work on the effects of incarceration on future criminality which is somewhat contradictory. For example, Aizer and Doyle (2015) and Mueller-Smith (2015) find that incarceration increases recidivism; while Mitchell, Cochran, Mears and Bales (2017) and Dobbie, Goldin and Yang (2018) find no effect; and Bhuller, Dahl, Løken and Mogstad (2020) suggest that prison time even reduces future criminality. In contrast, there is relatively little research on the impact of merely being convicted (not necessarily serving prison time) on future criminal behaviour. However, as with the work on incarceration, this research is also contradictory. For example, Ventura and Davis (2005) find that convictions reduce the likelihood of recidivism, while Chiricos, Barrick, Bales and Bontrager (2007) show the opposite effect.

As such, it is hard to appeal to the previous literature to defend the conditional mean independence assumption in our setting. Consequently, to test this assumption, we employ the popular approach of using the conviction tendency of quasi-randomly-assigned judges as an instrumental variable to uncover the causal effect of conviction on future criminality. To do this, a subsample of the primary dataset is taken for which we have confirmation that the assignment of judges to cases is quasi-random. Details of the sample used are provided in Section 3.2, and a full discussion of the regression analysis and results is given in Section 4.1. However, the results indicate that there is a small and insignificant effect of conviction on future criminality.

3.2 Sample and Variable Construction

Court case data from Virginia's 32 general district courts are used. The Virginia trial court system is broadly split between general district courts and circuit courts. General district courts make rulings on misdemeanours and infractions but only hold preliminary hearings on felony cases before transferring them to a circuit court. Circuit courts are the highest trial courts with general jurisdiction in Virginia; they hear more serious crimes and appeals from the general district courts.

The unit of observation is a single charge, and the data covers all arrests for which charges were filed for the years 2009-2018. Each observation provides information on the defendant's gender, race, and address, as well as details of the charge and the outcome of the criminal proceedings. The initial

dataset contains more than 20 million observations.

As discussed in Section 3.1, future criminality is used as the special regressor. This variable is constructed from several different measures of future criminal behaviour which must be individually calculated from the sample. To this end, a unique identifier for each individual based on their full name, gender, race, and day and month of birth is used to create the following variables: the number of arrests (including parole violations), the number of arrests for the same type of crime as the defendant is currently on trial,[10] the number of convictions, the number of convictions for the same type of crime, the dollar amount of fines charged, the number of days sentenced to prison, and the number of days for suspended sentences. Since data are only observed until the end of 2018, these measures of future criminal behaviour are averaged over the number of years between the date of the current trial (or the date of release if the defendant was convicted) and the end of the sample period.[11] We also construct these same seven measures for previous criminality using an analogous procedure. Having calculated these past and future variables, the first and last year of the data are dropped as 'burn-in periods' (losing 20% of observations).

At this stage, the sample is further restricted in several ways. Only those observations for which the defendant's race is categorised as black or white, respectively, are kept; there are too few observations on other race groups to obtain accurate estimates (6% loss of observations). We remove any individuals who are not residents of Virginia (13% loss of observations). This avoids the possibility that non-residents are treated differently to locals. Also, observations for which there is a guilty plea are removed (21% loss of observations). If a defendant pleads guilty, the decision is taken out of the hands of the court and the individual is convicted. If the defendant was guilty, a mistake would never be made, and if the defendant was innocent, a mistake would always be made. We are not interested in such scenarios in this paper, interesting as they no doubt are.

As discussed in Section 3.1, only infractions and misdemeanours are considered (11% loss of observations). This limits the severity of the potential punishment and thus mitigates the impact of the court decision on future criminal behaviour. Occasionally, a single trial will contain multiple charges against the defendant; this introduces a complex degree of dependence between observations. Recall that each observation represents a single charge. Thus, one trial may produce several observations in our data. To avoid issues of dependence, all observations for which the individual faces multiple charges (possibly multiple counts of the same charge) in a single trial are removed (23% loss

[10] The crime type is defined by the Virginia Crime Codes Statute Order. There are 1080 unique crime types in the sample.

[11] Unfortunately, we do not have information regarding parole, thus cannot adjust our measures to account for this.

of observations). Having made these restrictions, the sample contains 7.3 million observations.

Each of the seven aforementioned measures of future criminality is a viable choice for the special regressor. However, our method requires only a single variable. As such, we follow the increasingly popular path of using machine learning techniques to combine the measures of future criminality to create a single instrument which captures as much information as possible (see, for example, Lee, Lessler, and Stuart, 2010; Belloni, Chen, Chernozhukov and Hansen, 2012; Hartford, Lewis, Leyton-Brown and Taddy, 2017). In particular, gradient boosted regression trees are used.[12] Ideally, the outcome of interest in this first stage would be true guilt, but this is unobservable. Instead, we use conviction. Under the conditional mean independence assumption, the best linear predictor of guilt is also the best linear predictor of conviction. However, this is no longer true when using nonlinear prediction techniques, as we do. Nonetheless, the approach still yields a strong predictor for guilt, and evidence to this fact is given in Section 4.2. Recall that the goal is not to determine the true conditional mean of guilt, only to create a measure which is highly predictive of this guilt.

The data are randomly split into a training set (30%) and a hold-out set (70%). With the training data, 5-fold cross-validation and a grid search is used to find optimal choices for tuning parameters. Specifically, we choose the shrinkage parameter, the number of trees, the tree depth, the minimum number of observations in the terminal nodes, and the fraction of the sample randomly chosen to propose the next tree in the expansion. The optimal parameter values are then used to build a regression tree on the training dataset. At this stage, the training dataset is discarded from all proceeding analysis. Finally, the measure of future criminality is given as the predicted conviction probability from this regression tree on the hold-out data. The final sample size of this hold-out set is 5.1 million observations. In Appendix A, we provide a detailed discussion of the implications of using a special regressor constructed from a regression-type model, and how this can impact the likelihood of satisfying Assumption 3 or 3′.

Three further control variables are constructed from the data. First, two measures of crime severity are calculated using the average fine and average prison sentence for each type of crime (again, defined by the Virginia Crime Codes Statute Order with 1080 unique crimes in our sample). Specifically, we take a leave-one-out average of the fine charged for all cases where the defendant was found guilty for the same type of crime. An analogous variable is created for prison sentences. Secondly, to avoid the need for complicated nonparametric fixed-effects estimators, a ZIP code pseudo-fixed-effect is

[12] Other machine learning approaches could also be used. However, we found the highest out-of-sample correlation using this method.

calculated by taking a leave-one-out average of total arrests per-capita for each ZIP code. There are 902 ZIP code areas in Virginia with a mean population of 9325 and a median of 2940. We hope that including such a variable controls for some unobserved heterogeneity across neighbourhoods.

Indeed, Altonji and Mansfield (2018) give credibility to this idea. Translated into our context, they explain that if the decision to convict a defendant is based on both individual and neighbourhood characteristics and that individuals choose their neighbourhood endogenously, a bias can arise. However, under certain assumptions, controlling for means of observable individual factors at the neighbourhood level can "absorb all of the between-group variation in both observable and unobservable individual inputs" (pp. 2903). To achieve this perfect control of unobserved neighbourhood effects, the utility function of individuals choosing which neighbourhoods to live in must be additively separable in the amenities of the neighbourhood. Furthermore, the number of amenities which have an effect on court proceedings must not be larger than the number of neighbourhood averages included. That is, for full control of unobserved neighbourhood effects, we require judges to use a one-dimensional measure of neighbourhood quality in their decision to convict or acquit.

The choice to control for average total arrests within each ZIP code rather than other observable characteristics stems from a desire to avoid controlling for possible biases in the court proceedings. For example, suppose a bias against black individuals exists, this results in neighbourhoods with a large black population having a high conviction rate per-capita. If this neighbourhood conviction rate is included as a control variable, one channel through which racism works is partially closed off. In other words, we should not include controls which are a function of the outcome. By including a measure of criminality which is independent of court rulings, this is avoided.

We close this section with a final remark. Since the sample consists of individuals who have been arrested and subsequently charged, the analysis is conducted conditional on this fact. That is, the following objects are estimated

$$\begin{aligned} \alpha_1(x) &= P\left[Y = 1 \middle| Y^* = 0, X = x, A = 1\right], \\ \alpha_2(x) &= P\left[Y = 0 \middle| Y^* = 1, X = x, A = 1\right], \end{aligned}$$

where A denotes whether the individual has been arrested and charged ($= 1$) or not ($= 0$). Throughout, the notational dependence on A is dropped for convenience; however, the distinction should not be forgotten. We estimate the likelihood of a defendant - who has already been charged - being wrongfully convicted or wrongfully acquitted, respectively. These estimates are likely to be very different

for a defendant relative to a random member of the public.

3.2.1 Instrumental Variable Analysis

To test the conditional mean independence assumption using the IV approach discussed in Section 3.1, a small subset of the data is used. In particular, we take a subsample of six Virginia circuit courts from our primary dataset, namely: Chesterfield, Chesapeake, Hampton, Henrico, Newport, and Portsmouth. It is only for these courts that we have confirmation from the respective county court clerks that the assignment of judges to cases is random, providing there is not a subsequent action arising from the initial filing of a case. For example, probation violations are normally assigned to the judge who heard the original case. As such, judge assignment is taken to be random conditional on the type of trial.

While the judge is randomly assigned to a case, the courthouse where the case is heard is typically determined by where the offence occurred. Thus, year $\times$ courthouse fixed-effects are also included to control for the possibility that some courthouses are more lenient on average. Since each courthouse has several judges, by including these fixed-effects, the instrument is effectively the leniency of each judge relative to the other judges in the same year and the same courthouse.

For this IV analysis, the dataset is further restricted to cases which reach the final trial. The identity of the judge is only known in the final trial; therefore, cases which were concluded prior to the final trial must be dropped from the analysis. Being limited to only final trial cases in six circuit courts severely reduces the sample size: we are left with 5 656 observations.

The measure of judge leniency is constructed as the leave-one-out residualised average conviction rate for each judge after controlling for the type of trial and year $\times$ courthouse fixed-effects; this follows the previous literature (see, for example, Dobbie, Goldin and Yang, 2018). In particular, the following linear regression is first estimated

$$C_{ijt} \quad = \quad \gamma\alpha_{jt} + \beta X_{ijt} + \epsilon_{ijt},$$

where C_{ijt} denotes whether case i heard at courthouse j in year t resulted in a conviction, α_{jt} represents year $\times$ courthouse fixed-effects, and X_{ijt} is a set of dummy variables capturing the type of trial. $\hat{\epsilon}_{ijt}$ is used to denote the residual conviction decision. The residual leniency measure for case i heard by

judge k is then constructed as

$$\eta_{ikt} = \frac{1}{n_k - 1} \sum_{l \neq i} \hat{\epsilon}_{ljt},$$

where the sum is taken over all cases heard by judge k (excluding case i), and n_k is the total number of cases heard by judge k. Summary statistics on this leniency measure are provided in Section 3.3.

3.3 Descriptive Statistics

Table 1 reports the mean of each variable used in the primary analysis for defendants who are convicted and for those who are acquitted, respectively. There is almost no racial or gender difference in those who are acquitted versus convicted. Unsurprisingly, males make up the majority of the sample, and given that the population of Virginia is 62% white (non-Hispanic), it is also unsurprising that the sample is predominantly white. It is interesting to note that 29% of judges in Virginia are black[13], corresponding almost exactly with the proportion of black defendants. Thus, it is equally likely that a black defendant faces a white judge, as it is a white defendant is tried by a black judge.

Table 1: Mean Values (Full Sample)

	Convicted	Acquitted
White	0.71	0.69
Male	0.60	0.58
Future Criminality	0.81	0.79
Previous Arrests	1.46	1.51
Neighbourhood-Effect	0.64	0.67
Infraction	0.84	0.63
Crime Severity (Prison)	6.69	19.5
Crime Severity (Fine)	78.8	75.6
Observations	4,125,691	986,304

Notes: This table displays the means of the variables listed for the final sample of defendants from Virginia (selected according to the criteria given in Section 3.2) used to estimate the misclassification rates presented in Section 5.

Future criminality is only slightly higher for those who are convicted relative to those who are acquitted. In addition, there appears to be little difference in the number of previous arrests or the

[13] American Constitution Society for Law and Policy.

neighbourhood-effect across convicted and acquitted defendants. This is likely a reflection of the types of crimes which lead to a conviction. A higher fraction of crimes resulting in conviction are infractions rather than misdemeanours. This also explains why crime severity (measured by prison sentence) is lower for convicted crimes. In contrast, the severity of crime variable measured by the fine amount is similar across the two groups because infractions and misdemeanours typically have similar fine amounts, despite having different prison sentences.

3.3.1 Descriptive Statistics - Instrumental Variable Analysis

Table 2 presents mean values for the variables used in our IV analysis, again, separated by conviction status. The racial and gender differences for those convicted and acquitted, respectively, are small. However, the racial makeup of this subsample is quite different from that of the full sample. The reason lies in the severity of crimes heard by circuit courts as opposed to general district courts; recall that the data used for the IV analysis comes from six circuit courts. On average, blacks are on trial for more serious crimes. In the full sample, the fraction of blacks facing a misdemeanour charge is 27%, compared to only 17% for whites; the average prison sentence for cases against blacks is 13 days, versus 7.5 days for whites. Hence, blacks are disproportionately represented in circuit courts. This is also seen in the larger means of the crime severity variables and the smaller proportion of infractions in comparison to Table 1.

This racial difference may pose a concern for using the IV results of this subsample to extrapolate to the full sample. However, it seems reasonable to assume that if convictions for more serious crimes do not impact future criminality, then convictions for lesser crimes should also not affect future criminality. Nonetheless, perhaps convictions affect blacks differently to whites, and caution should still be applied in generalising our findings to the full sample. To alleviate concerns of this nature, the entire IV analysis is also conducted separately for blacks and whites. The results from this analysis are contained in Appendix B and are consistent with the baseline findings, suggesting that the difference in racial makeup between the two samples is not a concern.

Table 2: Mean Values (IV Subsample)

	Convicted	Acquitted
White	0.34	0.38
Male	0.66	0.63
Future Criminality	0.79	0.77
Previous Convictions	1.53	1.25
Previous Convictions (Same Crime)	0.32	0.21
Previous Arrests	2.27	1.97
Previous Arrests (Same Crime)	0.36	0.24
Previous Prison Time	27.9	21.5
Previous Suspended Prison Time	145	144
Previous Total Fines	96.4	81.5
Infraction	0.39	0.27
Crime Severity (Prison)	52.5	64.2
Crime Severity (Fine)	101	81.7
Observations	4,562	1,094

Notes: This table displays the means of the variables listed for the sample of defendants from six circuit county courts in Virginia (selected according to the criteria given in Section 3.2) used in the IV analysis to test the conditional mean independence restriction.

For the construction of the residualised judge leniency measure, there are 28 unique judges with an average judge hearing 202 cases. The leniency measure ranges from -0.17 to 0.22 with a standard deviation of 0.07. Moving from a judge at the 25% quantile to the 75% quantile increases the probability of conviction by 10.6 percentage points. Note that the average conviction rate in the subsample used for the IV analysis is 80.6%. The estimated distribution of residual judge leniency is given in Figure B.1 in Appendix B.

4 Research Design

4.1 Validity of the Conditional Mean Independence Assumption

In this section, we test whether future criminality satisfies the conditional mean independence condition. Recall that this condition is given by

$$E\left[Y - Y^* | Y^*, X, V\right] = E\left[Y - Y^* | Y^*, X\right],$$

or equivalently

$$E\left[Y | Y^*, X, V\right] = E\left[Y | Y^*, X\right].$$

In Section 3.1, it was argued that the most likely cause of a failure of this assumption is through conviction status affecting future criminality. To uncover the causal effect of conviction on future criminality, we use the leniency of quasi-randomly-assigned judges as an instrumental variable. Full details of the sample and the variables used are given in Section 3.

Table 3 presents results for the first-stage of the IV regression: a linear probability model of conviction status on residualised judge leniency. Recall that residualised judge leniency is the leave-one-out average conviction rate of a judge after controlling for the type of trial and courthouse×year fixed-effects. Column (1) reports the effect with no controls, column (2) adds the race, gender, and measures of previous criminality of the defendant. These measures include arrests, arrests for the same type of crime as they are currently on trial for, convictions, convictions for the same type of crime, fines charged, prison time, and suspended prison time. Column (3) additionally controls for case characteristics, including whether the crime is an infraction or misdemeanour and the severity of the crime (measured in terms of the average fine and prison sentence, respectively). Column (4) adds ZIP code fixed-effects. Throughout this section, all continuous regressors are standardised to have zero mean and unit variance, and standard errors for the estimated coefficients (clustered at the defendant level) are reported in parentheses.

Table 3: First Stage Regression

	Dependent variable:			
	Convicted			
	(1)	(2)	(3)	(4)
Judge Leniency	0.07	0.07	0.07	0.07
	(0.01)	(0.01)	(0.01)	(0.01)
White		−0.02	−0.02	−0.01
		(0.01)	(0.01)	(0.01)
Male		0.06	0.05	0.05
		(0.01)	(0.01)	(0.01)
Previous Criminality		✓	✓	✓
Case Characteristics			✓	✓
ZIP Code Fixed-Effects				✓
Observations	5,656	5,656	5,656	5,656
Adjusted R^2	0.03	0.04	0.07	0.08

Notes: This table reports results from first stage regressions using the subsample of six circuit courts from Virginia as detailed in Section 3.2. The dependent variable is a binary indicator for whether the case resulted in a conviction. Judge Leniency is the residualised leave-one-out average conviction rate of the judge after controlling for the type of trial and courthouse×year fixed-effects and is standardised to have unit variance. The other regressors are constructed as per the discussion in 3.2. Column (1) gives the simple regression of conviction status on judge leniency. Column (2) adds the defendant's race, gender, and seven measures of previous criminality. Column (2) includes whether the crime is an infraction or misdemeanour and the severity of the crime (measured by average fine and average prison sentence, respectively). Column (4) adds ZIP code fixed-effects. Standard errors for the estimated coefficients are reported in parentheses and are clustered at the individual level. * indicates significance at 10%, ** indicates significance at 5%, and *** indicates significance at 1%.

Across all four regressions, judge leniency has a highly significant effect. In particular, a one standard deviation increase in the judge leniency measure increases the probability of conviction by seven percentage points. It is also promising to see that this effect is constant irrespective of the control variables included; this provides good evidence of the quasi-random-assignment of judges to cases.

Nevertheless, we formally test this randomisation in Table 4. Here, the judge leniency measure is regressed on all case and defendant characteristics and ZIP code fixed-effects. The p-value for the joint significance of this regression is 0.2, suggesting that the leniency of the judge is unrelated to the case or the defendant. This adds further weight to the validity of the exclusion restriction.

Table 4: Test of Randomisation

	Dependent variable:
	Judge Leniency
White	−0.05
	(0.03)
Male	0.02
	(0.03)
Previous Criminality	✓
Case Characteristics	✓
ZIP Code Fixed-Effects	✓
Observations	5,656
Adjusted R^2	0.01
F Statistic	1.08 (df = 253; 5402)

Notes: This table reports results from a test of the randomisation of judge leniency using the subsample of six circuit courts from Virginia as detailed in Section 3.2. The dependent variable is judge leniency calculated as the residualised leave-one-out average conviction rate of the judge after controlling for the type of trial and courthouse×year fixed-effects and is standardised to have unit variance. The regressors are constructed as per the discussion in 3.2. Previous criminality includes the seven measures of previous criminality given in Section 3.2. Case characteristics include whether the crime is an infraction or misdemeanour and the severity of the crime (measured by average fine and average prison sentence, respectively). Standard errors for the estimated coefficients are reported in parentheses and are clustered at the individual level. The p-value for the F-test for the joint significance of the whole regression is 0.198. * indicates significance at 10%, ** indicates significance at 5%, and *** indicates significance at 1%.

Finally, we also check the validity of the monotonicity assumption for IV regressions. In Appendix B, Figure B.2 plots a univariate nonparametric version of the first stage regression of judge leniency on conviction. We see the probability of conviction is monotonically increasing in the measure of judge leniency and is approximately linear.

Table 5 contains the final IV results. In all four regressions, conviction does not have a significant effect on future criminality. However, this seems to be driven primarily by the large standard errors of the IV estimate. Nonetheless, the point estimate from the regression with the full set of controls indicates that the effect of being convicted increases the level of future criminality by 0.08 of a standard deviation - a small effect. Again, it is encouraging to see that the effect is relatively stable across the regressions.

The large standard errors corresponding to the effect of conviction on future criminality in Table 5 is a result of the relatively small sample used for this analysis. Thus, caution should be applied when drawing strong conclusions from these results since IV estimates can suffer from bias in small

samples. As a result, OLS estimates are also reported in Table B.1 in Appendix B. Interestingly, with the full set of controls, the effect is still small (0.03 of a standard deviation) and insignificant.

Table 5: IV Regression

	Dependent variable:			
	Future Criminality			
	(1)	(2)	(3)	(4)
Convicted	0.07	0.05	0.10	0.08
	(0.22)	(0.22)	(0.22)	(0.22)
White		0.25^{***}	0.28^{***}	0.24^{***}
		(0.03)	(0.03)	(0.03)
Male		0.29^{***}	0.25^{***}	0.27^{***}
		(0.03)	(0.03)	(0.03)
Previous Criminality		✓	✓	✓
Case Characteristics			✓	✓
ZIP Code Fixed-Effects				✓
Observations	5,656	5,656	5,656	5,656

Notes: This table reports results from four IV regressions using the subsample of six circuit courts from Virginia as detailed in Section 3.2. The dependent variable is future criminality calculated using the procedure given in Section 3.1 and is standardised to have unit variance. Conviction is a binary indicator for whether the defendant was convicted, it is instrumented by judge leniency. Judge Leniency is the residualised leave-one-out average conviction rate of the judge after controlling for the type of trial and courthouse×year fixed-effects. The other regressors are constructed as per the discussion in 3.2. Column (1) gives the simple IV regression of future criminality on conviction status. Column (2) adds the defendant's race, gender, and seven measures of previous criminality. Column (2) includes whether the crime is an infraction or misdemeanour and the severity of the crime (measured by average fine and average prison sentence, respectively). Column (4) adds ZIP code fixed-effects. Standard errors for the estimated coefficients are reported in parentheses and are clustered at the individual level. * indicates significance at 10%, ** indicates significance at 5%, and *** indicates significance at 1%.

As a robustness check, as mentioned in Section 3.3, separate estimates for blacks and whites are also provided in Appendix B. The conclusions in all cases are similar to those of the full sample. Overall, these findings suggest that the effect of conviction on future criminality is likely to be small and, consequently, the conditional mean independence assumption is likely to hold, at least approximately.

4.2 Validity of the Large Support Assumption

This section verifies the validity of the large support assumption. Throughout this analysis, the following set of controls are used: the neighbourhood-effect, the number of previous arrests of the defendant, the race and gender of the defendant, whether the crime is an infraction or misdemeanour,

and the severity of the crime measured by the average fine and the average prison sentence, respectively. To allow for the greatest flexibility, we estimate each model separately for the four race-gender groups.

The choice to control only for previous arrests rather than including other measures of previous criminality is based on the desire to avoid including variables which can be influenced by bias in the court proceedings. That is, we do not wish to control for covariates which may be a function of the outcome.

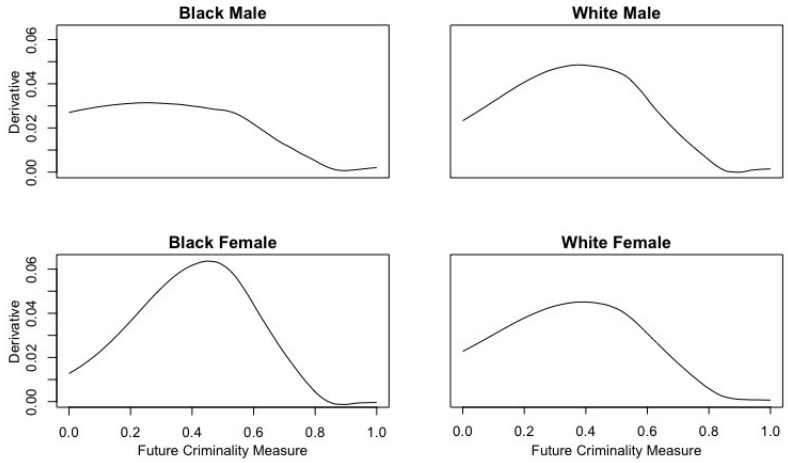
A local linear likelihood estimator with a logistic link function is used to estimate the nonparametric functions. Frölich (2006) showed in a series of Monte Carlo simulations that local likelihood logit estimation is substantially more precise than the Nadaraya-Watson estimator, the local linear kernel estimator, the semiparametric estimator of Klein and Spady (1993), and the parametric logit estimator in a binary choice model. A locally adaptive bandwidth is chosen using the intersection of confidence intervals (ICI) method (see Loader, 2006, for full details). The Epanechnikov kernel is used for all estimation procedures.

Our goal is to determine whether the large support condition is satisfied for given values of the set of regressors. Throughout, the neighbourhood-effect and the two measures of crime severity are set at their respective means, and the type of crime is fixed to be a misdemeanour when evaluating the estimators. We focus on the relationship between previous arrests and the type I and type II error, respectively. As such, we investigate the interval of values of the previous arrest measure for which the large support condition is satisfied.

Figure 4.1 gives representative plots of $(\partial/\partial v)\, E\left[Y|V=v, X=x\right]$ evaluated over the range of v (future criminality) with the previous arrests measure set at zero; this corresponds to no previous arrests within the sample period, i.e. the minimum value. It is clear from these plots that only the upper tail condition is satisfied. This indicates that there are individuals in the sample with high enough future criminality that their guilt can be perfectly predicted for the current crime. However, it is not possible to determine the underlying guilt of defendants with the lowest level of future criminality.

Nonetheless, in each case, the mode of $U|X=x$ is contained in the support of future criminality. Furthermore, the distributions appear relatively symmetric, giving hope to the validity of the mode-median coincidence condition (Assumption 4). Thus, $\alpha_2(x)$ is estimated using the procedure laid out in Section 2.1 and $\alpha_1(x)$ is estimated using the slightly more complex arguments of Section 2.3. These estimates are reported in Section 5.

Figure 4.1: Large Support Check (1)

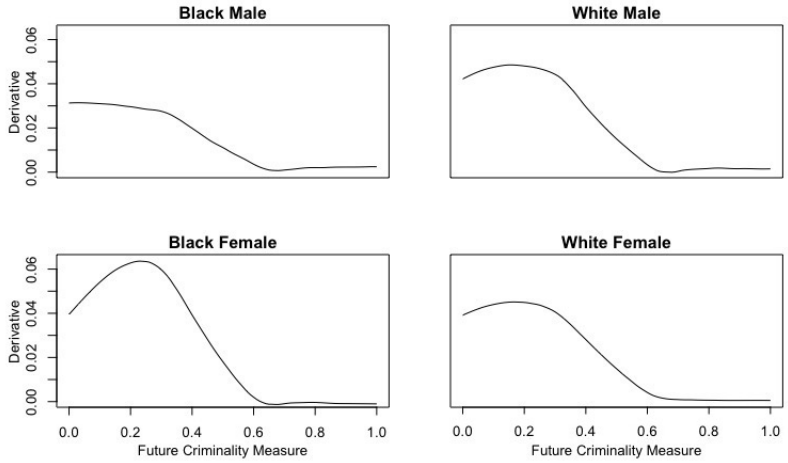


Notes: This figure plots estimates of $\partial E[Y|V=v, X=x]/\partial v$ over the range of v (future criminality) with previous arrests set at zero (corresponding to the minimum value). The previous arrests measure is calculated as the average number of previous arrests per annum since the start of the sample period and future criminality is constructed as outlined in Section 3.1. The crime type is set to be a misdemeanour, and the neighbourhood-effect variable and both measures of crime severity are set at their means. Each nonparametric function is estimated using a local linear logit estimator using the full dataset as detailed in Section 3.2 split into the respective race-gender groups.

Figure 4.2 displays analogous plots when the previous arrest measure is set to seven, i.e. seven arrests per annum since the start of the sample period. The choice of seven corresponds to the largest value of previous arrests for which the mode of $U|X=x$ is contained in the support $V|X=x$ for all race-gender groups. Note that the mode is deemed to be contained in the support if the maximum point is not at the boundary of the support.

Figures 4.1 and 4.2 highlight the tradeoff between the tail and the mode conditions; a larger value of previous arrests increases the chance of satisfying the tail constraint but at the cost of potentially losing the mode from the support. The intuition is that it is more difficult to predict those who are truly innocent when they already have a bad previous criminal record. Conversely, it is easier to predict whether a defendant is truly guilty if they already have a high number of prior arrests. As a result, in our context, it is not possible to estimate both misclassification rates for all possible covariate values. However, the upper tail condition is satisfied for all values of previous arrests. Thus, the probability that a guilty defendant is acquitted, $\alpha_2(\cdot)$, can be estimated over the full range of previous arrests.

Figure 4.2: Large Support Check (2)

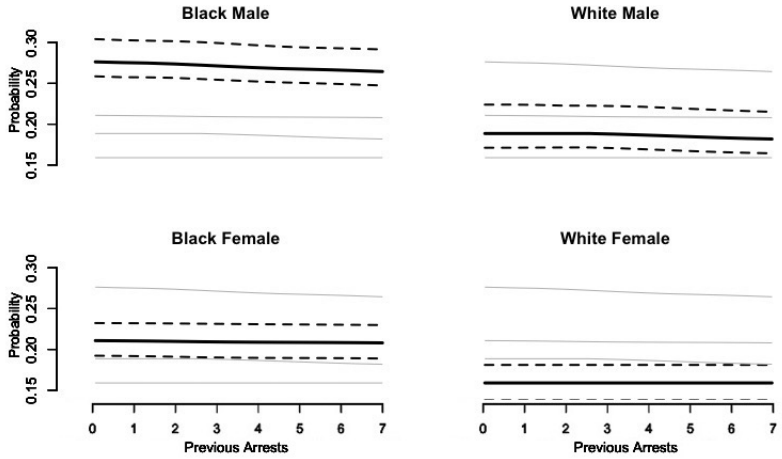


Notes: This figure plots estimates of $\partial E[Y|V=v, X=x]/\partial v$ over the range of v (future criminality) with previous arrests set at seven (corresponding to the largest value for which the mode of $U|X=x$ is contained in the support $V|X=x$ for all race-gender groups). The previous arrests measure is calculated as the average number of previous arrests per annum since the start of the sample period and future criminality is constructed as outlined in Section 3.1. The crime type is set to be a misdemeanour, and the neighbourhood-effect variable and both measures of crime severity are set at their means. Each nonparametric function is estimated using a local linear logit estimator using the full dataset as detailed in Section 3.2 split into the respective race-gender groups.

5 Results

In Figure 5.1, each panel plots the respective race-gender group's probability of being incorrectly convicted after being arrested as a function of previous arrests (displayed in bold black) together with a pointwise 95% confidence band based on a nonparametric bootstrap (displayed as dashed black lines). Plots of the other race-gender groups' probabilities are also included in each panel for comparison (displayed in grey). It should be noted that a theoretical justification for this bootstrap procedure is not provided; issues of inference are left for future work. The nonparametric functions themselves are estimated using a local linear logit estimator based on the identification scheme given in Section 2.3. The bandwidth is adaptive and chosen using the ICI method (see Loader, 2006, for details). The other control variables include the neighbourhood-effect, whether the crime is a misdemeanour or infraction, and the severity of the crime measured by the average fine and prison sentence, respectively (all variables are described in full in Section 3.2). Continuous control variables are set at their mean, and the crime is set to a misdemeanour (similar results were obtained when setting the crime to be an infraction).

Figure 5.1: Probability of Convicting an Innocent Defendant



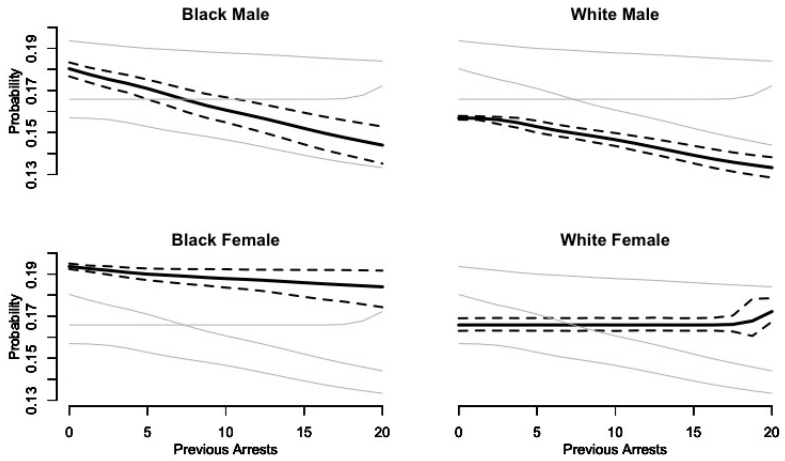
Notes: This figure plots estimates for the probability of convicting an innocent defendant using the identification scheme in Section 2.3. The nonparametric function is estimated using a local linear logit estimator with an adaptive bandwidth chosen using the 'ICI' method on the final sample of defendants (separated by race and gender) as detailed in Section 3.2. Previous arrests are calculated as the average number of arrests per annum. Other control variables are the neighbourhood-effect (defined in Section 3.2), whether the crime is a misdemeanour or infraction, and the severity of the crime measured by the average fine and the average prison sentence, respectively. Continuous variables are set at their mean and the crime as a misdemeanour. Each plot displays the respective estimate (solid black), pointwise 95% confidence band constructed using a nonparametric bootstrap (dashed black), and estimates of the other three race-gender groups (grey).

First, the likelihood of convicting an innocent black defendant is higher than that for an innocent white defendant, irrespective of gender. In particular, an innocent black male who has been arrested faces a worryingly high probability of being convicted. Indeed, each demographic group has a high chance of conviction when innocent. However, recall that the analysis is conducted conditional on being arrested for the crime. In order for an individual to be arrested, there must be compelling evidence against them. Thus, these estimates are likely to be quite different from estimates of the probability of convicting an innocent member of the public. However, the results are no less informative of the justice system; they merely isolate biases in the judicial process after arrest.

A bias against males relative to females also exists which is consistent across race groups. At the mean of previous arrests, innocent white males face a conviction probability of 18% in comparison to 16% for white females. Equally, innocent black males have a 28% probability of conviction in contrast to 21% for black females. Notice also that previous arrests have little effect on conviction probability; indeed, the 95% confidence band for each group contains a constant effect. As explained in Section 4.2, the probability of a false conviction cannot be identified when previous arrests are above seven due to a failure of the mode condition for black males. Thus, perhaps the range of average previous arrests per annum is too small to see a large effect.

Figure 5.2 plots analogous estimates for the probability of acquitting a guilty defendant. Importantly, since the right tail condition is satisfied - as shown in Section 4.2 - these estimates are calculated using the identification strategy of Section 2.1.

Figure 5.2: Probability of Acquitting a Guilty Defendant



Notes: This figure plots estimates for the probability of acquitting a guilty defendant using the identification scheme in Section 2.1. The nonparametric function is estimated using a local linear logit estimator with an adaptive bandwidth chosen using the 'ICI' method on the final sample of defendants (separated by race and gender) as detailed in Section 3.2. Previous arrests are calculated as the average number of arrests per annum. Other control variables are the neighbourhood-effect (defined in Section 3.2), whether the crime is a misdemeanour or infraction, and the severity of the crime measured by the average fine and the average prison sentence, respectively. Continuous variables are set at their mean and the crime as a misdemeanour. Each plot displays the respective estimate (solid black), pointwise 95% confidence band constructed using a nonparametric bootstrap (dashed black), and estimates of the other three race-gender groups (grey).

Again, there is a bias against males in favour of females, although the difference is relatively small. At the average value of previous arrests, the probability of acquittal for a guilty black male is 18%, compared to 19% for black females. Equally, guilty white males face a 16% chance of being acquitted, in comparison to a 17% likelihood for white females. Taken together with the results of Figure 5.1, it appears that the threshold for convicting a woman is lower than that for a man.

Figure 5.2 also shows that the effect of previous arrests on the probability of acquittal is more pronounced for men than women, and blacks relative to whites. In particular, across the range of previous arrests (from zero prior arrests to 20 per year), the acquittal probability for guilty black males falls from 18% to 15%, while for guilty white males there is a drop from 16% to 14%. Black females see their probability fall from 19% to 18%, but for white females it stays constant at 17%. It should not be surprising that the probability of wrongful acquittal falls as the number of previous arrests increases; this reflects the greater likelihood of convicting a defendant if they have a particularly criminal past.

With respect to the variance of our estimates, the confidence band is much narrower for the

probability of wrongful acquittal relative to that for wrongful conviction. This follows from the probability of wrongful conviction being estimated using the more complex identification scheme of Section 2.3 because the left tail condition is not satisfied. Figure 5.2 also shows that the confidence band widens for larger values of previous arrests; this is to be expected since the data is more sparse in the upper tail of previous arrests.

Interestingly, despite the probability of wrongful conviction being higher for blacks than whites, the probability of wrongful *acquittal* is, in general, higher for blacks. Furthermore, this is consistent across genders. It is difficult to pinpoint precisely why such a pattern exists; however, it is constructive to discuss possible causes.

One potential explanation for this finding relates to the adequacy of the model. It may be that while the conditional mean independence assumption holds over the whole sample, it may not hold individually for whites. If conviction has a positive effect on future criminality for whites, this would result in an under-estimation of the probability of wrongful acquittal for whites. However, Tables B.2 and B.3 in Appendix B refute such a premise. These tables display results of the IV regression discussed in Section 4.1 estimated on the subsample of black and white defendants, respectively. The results are similar in each case and suggest the conditional mean independence assumption holds for both races.

This contradictory result could instead be explained by black defendants having more procedural flaws in their cases than white defendants. For example, if officers violate the constitutional rights of blacks more often than whites, such as through illegal searches, the judge is obliged to dismiss more cases against blacks despite perhaps believing the defendant to be guilty. Thus, although our method detects incorrect decisions, it may not necessarily be a result of a judicial error, but rather a policing error.

In the following section, we delve deeper into this idea and attempt to shed light on the wider judicial system by developing a simple theoretical model where the parameters are calibrated based on the empirical findings.

5.1 Calibration-Based Model of Discrimination

Consider the following stylised model. In the first stage, for each individual who comes into contact with the police regarding a given crime, the police draw a noisy signal of the suspect's guilt, denoted $e_1(G)$, where G denotes whether the suspect is guilty ($= 1$) or innocent ($= 0$). For simplicity, assume this 'strength-of-evidence' measure is distributed as $e_1(G) \sim N(\mu_G, 1)$, where we normalise $\mu_0 = 0$

and assume $\mu_1 > 0$. Police make arrests according to the following threshold decision rule

$$A = \begin{cases} 1 & \text{if } e_1(G) \geq T_P(R) \\ 0 & \text{if } e_1(G) < T_P(R), \end{cases}$$

where A denotes an arrest ($= 1$) or no arrest ($= 0$), and $T_P(R)$ denotes the police decision threshold for $R \in \{Black, White\}$.

For each arrested defendant, a second strength-of-evidence variable is drawn, $e_2(G)$, also from $N(\mu_G, 1)$, which is only observed after the arrest. This captures the fact that a substantial amount of evidence is likely to be collected post-arrest (this also includes that the case put forward by the prosecution is constructed post-arrest and can be considered as part of the strength-of-evidence). Note that to keep the model tractable, it is assumed that the variance of the signal before and after arrest for both guilty and innocent suspects is the same. Although this is unlikely to be the case in reality, it prevents the proliferation of parameters and allows the identification of more important model characteristics.

Similarly to the police, the judge convicts the defendant based on the following threshold decision rule

$$C = \begin{cases} 1 & \text{if } e(G) \geq T_J(R) \\ 0 & \text{if } e(G) < T_J(R), \end{cases}$$

where $e(G) = \frac{1}{2}(e_1(G) + e_2(G))$, C denotes whether the judge convicts ($= 1$) or acquits ($= 0$) the defendant, and $T_J(R)$ is the judge decision threshold. The form of $e(G)$ can be interpreted as the judge taking an equally weighted average of both signals, that is, evidence collected pre and post-arrest. We go on to check the sensitivity of our results to this assumption below.

Based on this model, the type I error for a given race category R is

$$\begin{aligned} P[C = 1 | G = 0, A = 1] &= P\left[e(0) \geq T_J(R) \middle| e_1(0) \geq T_P(R)\right] \\ &= \frac{P[Z_1 > T_J(R), Z_2 > T_P(R)]}{1 - \Phi(T_P(R))} \end{aligned} \tag{12}$$

where (Z_1, Z_2) are bivariate standard normal with correlation equal to $\sqrt{1/2}$. In a similar fashion,

the type II error in this model is written as

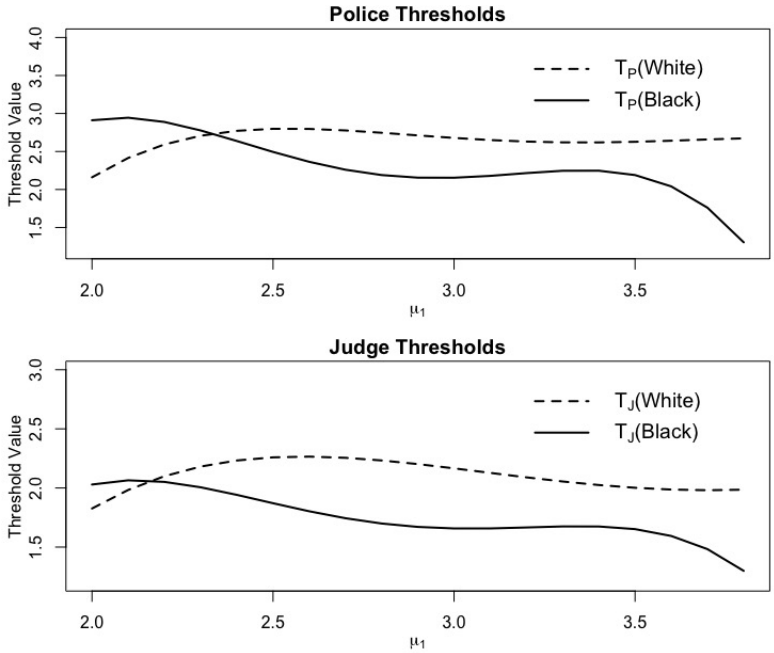
$$\begin{aligned} P\left[C=0|G=1,A=1\right] &= P\left[e\left(1\right)\leq T_J\left(R\right)\left|e_1\left(1\right)\geq T_P\left(R\right)\right.\right] \\ &= 1-\frac{P\left[Z_1>T_J\left(R\right)-\mu_1, Z_2>T_P\left(R\right)-\mu_1\right]}{1-\Phi\left(T_P\left(R\right)-\mu_1\right)}. \end{aligned} \tag{13}$$

These error probabilities do not admit a closed-form solution. However, it is possible to simulate and calibrate this theoretical model (for a given μ_1) using the average empirical estimates for the type I and type II probabilities for each race group to solve for the values of $T_P\left(Black\right)$, $T_J\left(Black\right)$, $T_P\left(White\right)$, and $T_J\left(White\right)$. Figure 5.3 plots the resulting threshold values for both the police and the judge for both blacks and whites across a range of plausible values for μ_1. Recall that $\mu_0 = 0$, so a choice of $\mu_1 = 2$ involves considerable overlap between the two distributions (approximately 32%), while $\mu_2 = 4$ leads to less than 5% overlap between the two distributions.

Figure 5.3 shows that the threshold for both arresting and convicting whites is generally higher than for blacks. Thus, although our empirical results seem contradictory at first sight, they are largely consistent with a model where there is discrimination against blacks by both police officers and judges. Only when the distribution of the strength-of-evidence for guilty defendants is very close to the distribution for innocent defendants, do the results show a conflicting story.

Note that in this model, the judge places equal weight on the strength-of-evidence signal drawn pre-arrest as the signal drawn post-arrest. As a result, this model implicitly imposes that the correlation between the signal received by the police, $e_1\left(G\right)$, and the signal received by the judge, $e\left(G\right)$ is equal to $\sqrt{1/2} \approx 0.71$. However, it may be that the judge places more weight on post-arrest evidence than pre-arrest evidence, or vice versa. In such situations, the correlation between $e_1\left(G\right)$ and $e\left(G\right)$ will change. As a robustness check, we report results for correlations of 0.44 (equivalent to the judge placing twice as much weight on post-arrest evidence) and 0.89 (equivalent to the judge placing twice as much weight on pre-arrest evidence) in Appendix B. In each case, the results are qualitatively similar to the results given in Figure 5.3.

Figure 5.3: Calibrated Threshold Values



Notes: This figure plots the estimated police and judge decision thresholds from the model outlined in Section 5.1 for white and black defendants, respectively. For each race group, the estimated thresholds are obtained by equating the type I and type II error probabilities given in equations (12) and (13) with their respective empirical estimates and solving for the values of $T_P(Black), T_J(Black), T_P(White), T_J(White)$. In particular, the average empirical estimates of the type I error are 0.25 and 0.17 for blacks and whites, respectively, and 0.18 and 0.16 for the average empirical type II error for blacks and whites, respectively. The correlation between the strength-of-evidence signal seen by the police and the signal seen by the judge is $\sqrt{1/2}$.

6 Conclusion

In this paper, we estimate the likelihood of both wrongful conviction and wrongful acquittal using data on more than five million court cases from Virginia. Our method is based on reframing the problem in the context of misclassified binary choice models where the misclassification rates can be interpreted as type I and type II errors, respectively. We give new nonparametric identification results for these models that admit simple estimators and which are likely to be of independent interest. We also provide methods to test the identifying assumptions and give alternative estimation schemes for cases which fail these tests. In our empirical context, a thorough discussion of the identification conditions is provided along with evidence of their validity. This includes an analysis of the effect of conviction on future criminality using the quasi-random-assignment of judges to cases as an instrumental variable.

We find that blacks, relative to whites, have a higher probability of conviction when guilty but also have a higher probability of acquittal when innocent. However, we go on to show that such a result,

although seemingly contradictory, is, in fact, consistent with a theoretical model where the threshold for evidence to arrest blacks is lower than for whites and the threshold for evidence to convict blacks is also lower than for whites. In addition, our results also reveal that males face both a higher probability of conviction when innocent and a lower probability of acquittal when guilty relative to females.

There is still further work needed in this area that is beyond the scope of this paper. Most notably, we do not provide inference procedures for our estimator. Bootstrap confidence bands are given in the empirical analysis; however, no theoretical justification is provided. It would also be worthwhile to develop a formal test of the large support condition based on our heuristic arguments. Finally, due to data limitations, we are silent regarding the performance of judges relative to juries. It would be of great interest to explore if - and when - one type of trial is less prone to error than the other.

References

[1] Acker, J.R. (2017) Taking stock of innocence: Movements, mountains, and wrongful convictions. *Journal of Contemporary Criminal Justice*. 33(1), pp. 8–25.

[2] Aizer, A. and J.J. Doyle Jr (2015) Juvenile incarceration, human capital, and future crime: Evidence from randomly assigned judges. *Quarterly Journal of Economics*. 130(2), pp. 759-803.

[3] Altonji, J.G. and R.K. Mansfield (2018) Estimating group effects using averages of observables to control for sorting on unobservables: School and neighborhood effects. *American Economic Review*. 108(10), pp. 2902-46.

[4] Belloni, A., Chen, D., Chernozhukov, V. and C. Hansen (2012) Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*. 80(6), pp. 2369-2429.

[5] Berry, S.T. and P.A. Haile (2014) Identification in differentiated products markets using market level data. *Econometrica*. 82(5), pp. 1749-1797.

[6] Bhuller, M., Dahl, G.B., Løken, K.V. and Mogstad, M. (2020) Incarceration, recidivism, and employment. *Journal of Political Economy*. 128(4), pp.1269-1324.

[7] Bjerk, D. and E. Helland (2019) What can DNA exonerations tell us about racial differences in wrongful conviction rates? *Journal of Law and Economics* (Forthcoming).

[8] Chiricos, T., Barrick, K., Bales, W. and S. Bontrager (2007) The labeling of convicted felons and its consequences for recidivism. *Criminology.* 45(3), pp. 547-581.

[9] Dobbie, W., Goldin, J. and C.S. Yang (2018) The effects of pretrial detention on conviction, future crime, and employment: Evidence from randomly assigned judges. *American Economic Review*. 108(2), pp. 201-40.

[10] Frölich, M. (2006) Non-parametric regression for binary dependent variables. *The Econometrics Journal*. 9(3), pp. 511-540.

[11] Goh, C. (2018) Rate-optimal estimation of the intercept in a semiparametric sample-selection model. *Econometric Theory* (Forthcoming).

[12] Gross, S.R. and B. O'Brien (2008) Frequency and predictors of false conviction: Why we know so little, and new data on capital cases. *Journal of Empirical Legal Studies*, 5(4), pp. 927-962.

[13] Gross, S.R., O'Brien, B., Hu, C. and E.H. Kennedy (2014) Rate of false conviction of criminal defendants who are sentenced to death. *Proceedings of the National Academy of Sciences*. 111(20), pp. 7230-7235.

[14] Hartford, J., Lewis, G., Leyton-Brown, K. and M. Taddy (2017) Deep IV: A flexible approach for counterfactual prediction. *Proceedings of the 34th International Conference on Machine Learning*. 70, pp. 1414-1423.

[15] Hausman, J.A., Abrevaya, J. and F.M. Scott-Morton (1998) Misclassification of the dependent variable in a discrete-response setting. *Journal of Econometrics*. 87, pp. 239-269.

[16] Heckman, J. J. and S. Navarro (2007) Dynamic discrete choice and dynamic treatment effects. *Journal of Econometrics*. 136, pp. 341-396.

[17] Khan, S. and D. Nekipelov (2018) Information structure and statistical information in discrete response models. *Quantitative Economics*. 9(2), pp.995-1017.

[18] Klein, R.W. and R.H. Spady (1993) An efficient semiparametric estimator for binary response models. *Econometrica*. pp. 387-421.

[19] Lee, B.K., Lessler, J. and E. A. Stuart (2010) Improving propensity score weighting using machine learning. *Statistics in Medicine*. 29(3), pp. 337-346.

[20] Lewbel, A. (1998) Semiparametric latent variable model estimation with endogenous or mismeasured regressors. *Econometrica*. pp. 105-121.

[21] Lewbel, A. (2000a) Semiparametric qualitative response model estimation with unknown heteroscedasticity or instrumental variables. *Journal of Econometrics*. 97(1), pp. 145-177.

[22] Lewbel, A. (2000b) Identification of the binary choice model with misclassification. *Econometric Theory*. 16(4), pp. 603-609.

[23] Lewbel, A. and X. Tang (2015) Identification and estimation of games with incomplete information using excluded regressors. *Journal of Econometrics*. 189(1), pp. 229-244.

[24] Loader, C. (2006) *Local Regression and Likelihood*. Springer Science & Business Media.

[25] Magnac, T. and E. Maurin (2007) Identification and information on monotone binary models. *Journal of Econometrics*. 139(1), pp. 76-104.

[26] Manski, C.F. (1988) Identification of binary response models. *Journal of the American Statistical Association*. 83(403), pp. 729-738.

[27] Mitchell, O., Cochran, J.C., Mears, D.P. and W.D. Bales (2017) Examining prison effects on recidivism: A regression discontinuity approach. *Justice Quarterly*. 34(4), pp. 571-596.

[28] Mueller-Smith, M. (2015) The criminal and labor market impacts of incarceration. Working Paper.

[29] Risinger, D.M. (2006) Innocents convicted: An empirical justified factual wrongful conviction rate. *Journal of Criminal Law and Criminology*. 97, pp. 761.

[30] Spencer, B.D. (2007) Estimating the accuracy of jury verdicts. *Journal of Empirical Legal Studies*. 4(2), pp. 305-329.

[31] Ventura, L.A. and G. Davis (2005) Domestic violence: Court case conviction and recidivism. *Violence Against Women*. 11(2), pp. 255-277.

[32] Zalman, M. (2017) Wrongful Convictions: A Comparative Perspective. *Journal of Contemporary Criminal Justice*. 33(1), pp. 1-7.

Appendix A

Discussion of Mode-Median Coincidence Restriction

In this section, a discussion of the mode-median coincidence restriction of Assumption 4 is provided. We begin with the following theorem:

Theorem 2 For each $x \in supp\,(X)$, let $c_x \in \mathbb{R}$ (for some arbitrary choice c_x) and let $G^*\,(\cdot|x)$ be a function: $[l_V^x, r_V^x] \to [0,1]$. Suppose the following conditions hold: (i) $G^*\,(\cdot|x)$ is non-decreasing and continuously differentiable on $[l_V^x, r_V^x]$;[14] (ii) $\frac{\partial}{\partial v} G^*\,(v|x)$ has a unique maximiser $\bar{v}(x)$ on $[l_V^x, r_V^x]$. Then, for any G^* satisfying (i) and (ii), there exists a pair $(h(x), F_{U|X}(u|x))$ such that a set of random variables (Y^*, X, V, U) satisfies Assumption 2, Mode $[U|X=x]$ satisfies

$$\text{Mode}\,[U|X=x] = c_x, \tag{14}$$

and

$$G^*\,(v|x) = P\,[Y^*=1|\,(X,V)=(x,v)]$$

for each $v \in [l_V^x, r_V^x]$ and each $x \in supp\,(X)$.

Note that we can set $c_x = 0$ since the choice is arbitrary; however, we consider a non-zero c_x when comparing the conditional mode restriction of equation (14) with Assumption 4 in the main text. Theorem 2 highlights the role of the conditional mode restriction as a location normalisation in monotone discrete choice models to identify $h(x)$ and $F_{U|X}\,(u|x)$. Given the monotonicity of the model, equation (14) does not impose any significant restriction on the functional form of the CPP, $G^*\,(v|x)$, except for the maximiser condition (ii) which is quite mild.

To compare equation (14) and Assumption 4, suppose that Y were observable and thus $G^*\,(v|x)$ is identifiable. Then, letting $c_x = \text{Median}\,[U|X=x]$ gives the restriction in Assumption 4: Mode $[U|X=x] =$ Median $[U|X=x]$. In this case, Assumption 4 would be testable since both Mode $[U|X=x]$ and Median $[U|X=x]$ could be separately identified as $\bar{v}(x)$ and the value v which satisfies $G^*(v|x) = 1/2$, respectively; thus, Assumption 4 could be easily rejected unless $G^*(\bar{v}(x)|x) = 1/2$. However, when Y^* is unobservable - as in our empirical context - Median$[U|X=x]$ is not identifiable. Therefore, Assumption 4 is not in general testable but imposes a restriction on the form of $G^*(v|x)$, the CPP.

Finally, if $P[Y^*=1|(X,V)=(x,v)]$ were identifiable, identification of $h(x)$ and $F_{U|X}(u|x)$ could

[14] We define $\frac{\partial}{\partial v} G^*\,(v|x)$ as the one-sided derivative at each end point of the support.

be established, since they are uniquely determined by $G^*(v|x)$ under (i) and (ii), as argued in the proof of Theorem 2.[15] Thus, this theorem can be seen as analogous to Magnac and Maurin's (2007) representation result which is stated under an orthogonality moment condition (corresponding to $E[UX] = 0$ in the present context). We close this section with the theorem's proof:

Proof of Theorem 2 Recall that, given Assumption 2, $P\left[Y^* | \left(X, V\right) = (x, v)\right] = F_{U|X}\left(v + h(x)|x\right)$. Thus, it is sufficient to show that for each $G^*\left(\cdot|x\right)$ which satisfies (i) and (ii), there exists some $(h(x), F_{U|X}(u|x))$ such that $\text{Mode}\left[U|X = x\right] = c_x$ and

$$G^*\left(v|x\right) = F_{U|X}\left(v + h(x)|x\right).$$

Let $\bar{v}(x) := \text{argmax}_{v \in [l_V^x, r_V^x]} \frac{\partial}{\partial v} G^*\left(v|x\right)$ and define $h(x) := c_x - \bar{v}(x)$. Given this $h(x)$, construct $F_{U|X}\left(\cdot|x\right)$ as $F_{U|X}\left(v + h(x)|x\right) := G^*\left(v|x\right)$ for each $v \in [l_V^x, r_V^x]$, or equivalently

$$F_{U|X}\left(u|x\right) := G^*\left(u - h(x)|x\right) \tag{15}$$

for each $u \in [l_V^x + h(x), r_V^x + h(x)]$. By construction, $F_{U|X}\left(u|x\right)$ is differentiable and at $u = c_x$,

$$\frac{\partial}{\partial v} F_{U|X}\left(c_x|x\right) = \frac{\partial}{\partial v} G^*\left(c_x - h(x)|x\right) = \frac{\partial}{\partial v} G^*\left(\bar{v}(x)|x\right).$$

Thus, if $G^*\left(l_V^x|x\right) = 0$ and $G^*\left(r_V^x|x\right) = 1$, we can check that the distribution of $U|X = x$ is fully specified by equation (15) and satisfies equation (14). Otherwise, we can appropriately define the support of $U|X = x$, and the values of $f_{U|X}\left(u|x\right) = \frac{\partial}{\partial u} F_{U|X}\left(u|x\right)$ for $u < l_V^x + h(x)$ or $u > r_V^x + h(x)$, so that $f_{U|X}(u|x) < \frac{\partial}{\partial v} F_{U|X}\left(\bar{v}(x)|x\right)$ for any $u \neq \bar{v}(x)$, $F_{U|X}\left(l_V^x + h(x)|x\right) = G^*\left(l_V^x|x\right)$, and $F_{U|X}\left(r_V^x + h(x)|x\right) = G^*\left(r_V^x|x\right)$. ■

Derivation of Integral Form for Limit Object

Here, we outline how to derive the integral form for the limit object as given in equation (11) in the main text. Note that for each $(x, \tilde{v})$,

$$P\left[Y = 1 | \left(X, V\right) = (x, \tilde{v})\right] = \alpha_1(x) + \left[1 - \alpha_1(x) - \alpha_2(x)\right] F_{U|X}\left(\tilde{v} + h(x)|x\right)$$

[15] Based on this identification result, a new non/semiparametric estimator for latent-variable binary choice models could be constructed, although this is not pursued in this paper. To the best of our knowledge, there has been no study that considers the conditional mode restriction as in equation (14) for such models.

and

$$\begin{aligned}\int_{\tilde{v}}^{r_V^x} (\partial/\partial v)\, P\left[Y=1|\,(X,V)=(x,v)\right] dv &= \left[1-\alpha_1(x)-\alpha_2(x)\right] \int_{\tilde{v}}^{r_V^x} f_{U|X}\left(\tilde{v}+h(x)|x\right) dv \\ &= \left[1-\alpha_1(x)-\alpha_2(x)\right]\left[1-F_{U|X}\left(\tilde{v}+h(x)|x\right)\right].\end{aligned}$$

These two equations imply

$$1-\alpha_2(x) = P\left[Y=1|\,(X,V)=(x,v)\right] + \int_{\tilde{v}}^{r_V^x} (\partial/\partial v)\, P\left[Y=1|\,(X,V)=(x,v)\right] dv.$$

This provides an alternative estimator for $\alpha_2(x)$. However, this equation is based on an arbitrary choice $\tilde{v}$. In the hope of providing a more robust estimation procedure, we take the expectation over $\tilde{v}$. That is,

$$\begin{aligned}& 1-\alpha_2(x) \\ = \;& P\left[Y=1|X=x\right] + \int_{l_V^x}^{r_V^x} \left[\int_{\tilde{v}}^{r_V^x} (\partial/\partial v)\, P\left[Y=1|\,(X,V)=(x,v)\right] dv\right] f_{V|X}\left(\tilde{v}|x\right) d\tilde{v}.\end{aligned}$$

Furthermore, by Fubini's theorem,

$$\begin{aligned}& \int_{l_V^x}^{r_V^x} \left[\int_{\tilde{v}}^{r_V^x} (\partial/\partial v)\, P\left[Y=1|\,(X,V)=(x,v)\right] dv\right] f_{V|X}\left(\tilde{v}|x\right) d\tilde{v} \\ = \;& \int_{l_V^x}^{r_V^x} \left[\int_{l_V^x}^{v} (\partial/\partial v)\, P\left[Y=1|\,(X,V)=(x,v)\right] f_{V|X}\left(\tilde{v}|x\right) d\tilde{v}\right] dv \\ = \;& \int_{l_V^x}^{r_V^x} (\partial/\partial v)\, P\left[Y=1|\,(X,V)=(x,v)\right] F_{V|X}\left(v|x\right) dv.\end{aligned}$$

Thus,

$$1-\alpha_2(x) = P\left[Y=1|X=x\right] + \int_{l_V^x}^{r_V^x} (\partial/\partial v)\, P\left[Y=1|\,(X,V)=(x,v)\right] F_{V|X}\left(v|x\right) dv,$$

and the result is obtained. ■

Regression-Based Construction of Special Regressor

In this section, we discuss the construction of the special regressor V introduced in Section 2.1. As detailed in Section 3.2, we use the future criminal behaviour of the defendant as the special regressor but since there are several measures of future criminality, we construct a scalar V from these measures.

Denote by W a vector of such criminality measures, where W may include discretely distributed components. Recall that one of the basic conditions for V is that it is continuously distributed (supposed in Assumption 2). While V could be simply defined as an average of the components of W, where the averaging may lead to a smoother distribution function (and thus at least approximate continuity), we construct V through a (machine-learning-type) regression of Y on (X, W), where Y is the observable but misclassified version of Y^*, and X is the observable characteristic vector introduced in Section 2.1. Here, we discuss what form the regression should take such that the resulting V is likely to satisfy Assumption 3 or $3'$.

First, consider the following nonparametric regression of Y on (X, W):

$$Y = \kappa(X, W) + \epsilon, \tag{16}$$

where $E[\epsilon|X, W] = 0$ and κ is the regression (conditional expectation) function. Note that $\kappa(X, W) \in [0, 1]$ since $Y \in \{0, 1\}$. In our empirical work, V is an estimated probability from a gradient boosted regression tree which lies in $[0, 1]$.

While we could define

$$V_i = \kappa(X_i, W_i), \tag{17}$$

for each individual i, we claim this is not likely to be a sensible choice when W has sufficient predictability for Y. To this end, suppose the following slightly strengthened version of Assumption 1:

Assumption 1′

$$E[Y|Y^*, X, W] = E[Y|Y^*, X] \text{ almost surely,}$$

which implies Assumption 1 since V is assumed to be defined as a function of (X, W). This allows the misclassification error to be written as

$$E[Y^* - Y|(X, W) = (x, w)] = -\alpha_1(x) + [\alpha_1(x) + \alpha_2(x)] E[Y^*|(X, W) = (x, w)],$$

which can be derived analogously to (3) in Section 2.1. Given (17), the law of iterated expectations leads to

$$E[Y^* - Y|(X, V) = (x, v)] = -\alpha_1(x) + [\alpha_1(x) + \alpha_2(x)] E[Y^*|(X, V) = (x, v)]. \tag{18}$$

Note, by the definition of κ, we can also write

$$E\left[Y^{*}|X,W\right]=\kappa\left(X,W\right)+E\left[Y^{*}-Y|X,W\right]. \tag{19}$$

Taking (19), (18), (17), and the law of iterated expectations, we can write

$$E\left[Y^{*}|\left(X,V\right)=\left(x,v\right)\right] \quad = \quad v-\alpha_{1}\left(x\right)+\left[\alpha_{1}\left(x\right)+\alpha_{2}\left(x\right)\right]E\left[Y^{*}|\left(X,V\right)=\left(x,v\right)\right]. \tag{20}$$

Now, suppose there exists some $\left(x,w\right)$ such that $\kappa\left(x,w\right)=0$, i.e. $Y=0$ can be perfectly predicted. Since $v=\kappa\left(x,w\right)$, and $E\left[Y^{*}|\left(X,V\right)=\left(x,v\right)\right]\geq 0$ we have

$$-\alpha_{1}\left(x\right)+\left[\alpha_{1}\left(x\right)+\alpha_{2}\left(x\right)\right]E\left[Y^{*}|\left(X,V\right)=\left(x,0\right)\right] \quad \geq \quad 0.$$

Therefore,

$$E\left[Y^{*}|\left(X,V\right)=\left(x,0\right)\right] \quad \geq \quad \frac{\alpha_{1}\left(x\right)}{\alpha_{1}\left(x\right)+\alpha_{2}\left(x\right)}, \tag{21}$$

provided that $0<\alpha_{1}\left(x\right)+\alpha_{2}\left(x\right)<1$.

On the other hand, suppose there exists some $\left(x,\tilde{w}\right)$ such that $\kappa\left(x,\tilde{w}\right)=1$. Then, using (20), we have

$$E\left[Y^{*}|\left(X,V\right)=\left(x,1\right)\right]=1-\alpha_{1}\left(x\right)+\left[\alpha_{1}\left(x\right)+\alpha_{2}\left(x\right)\right]E\left[Y^{*}|\left(X,V\right)=\left(x,1\right)\right].$$

Since $E\left[Y^{*}|\left(X,V\right)=\left(x,1\right)\right]\leq 1$,

$$-\alpha_{1}\left(x\right)+\left[\alpha_{1}\left(x\right)+\alpha_{2}\left(x\right)\right]E\left[Y^{*}|\left(X,V\right)=\left(x,1\right)\right]\leq 0.$$

Therefore,

$$E\left[Y^{*}|\left(X,V\right)=\left(x,1\right)\right] \quad \leq \quad \frac{\alpha_{1}\left(x\right)}{\alpha_{1}\left(x\right)+\alpha_{2}\left(x\right)}. \tag{22}$$

Degeneracy of V

From (21) and (22), $E\left[Y^{*}|\left(X,V\right)=\left(x,v\right)\right]$ is degenerated to $\alpha_{1}\left(x\right)/\left[\alpha_{1}\left(x\right)+\alpha_{2}\left(x\right)\right]$ when it is increasing in v (as supposed in Assumption 2). That is, if predictions from the regression in (16) are used for V, there would be no hope of satisfying the large support condition given Assumptions 1

and 2 and sufficient predictability of Y from W.[16] Thus, it is not sensible to use predictions from a regression of the form in (16) as the special regressor.

Alternative construction of V

Instead, our special regressor is based on a regression of the form

$$Y = \widetilde{\kappa}(X_1, W) + \tilde{\epsilon},$$

where $E\left[\tilde{\epsilon}|X_1, W\right] = 0$ and X_1 is a vector consisting of subcomponents in X, and we let $V = \widetilde{\kappa}(X_1, W)$.

For the above regression, the corresponding expression for (19) is now given by

$$E\left[Y^*|X_1, W\right] = \widetilde{\kappa}(X_1, W) + E\left[Y^* - Y|X_1, W\right].$$

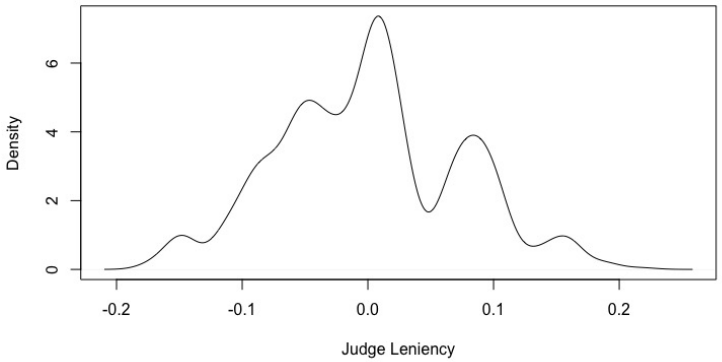
This cannot be (directly) combined with (20) and, consequently, does not result in counterparts to the inequalities in (21) and (22). Thus, the degeneracy result is avoided.

It is worthwhile to note that choosing an X_1 that is not sufficiently rich is important. This helps to ensure that $E[Y|Y^*, X_1, W] \neq E[Y|Y^*, X_1]$, which is required to avoid the degeneracy problem. For this reason, choosing $X_1 = \emptyset$ (i.e., $V = \widetilde{\kappa}(W)$ without X_1) appears to be a sensible choice, and is the choice we make in our empirical setting.

[16] The case of $v = 0$ (resp. $v = 1$) corresponds to the violation of the lower (resp. upper) tail condition of Assumption 3.

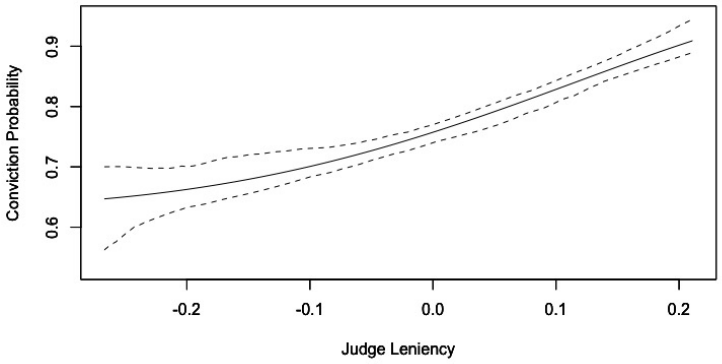
Appendix B

Figure B.1: Distribution of Judge Leniency



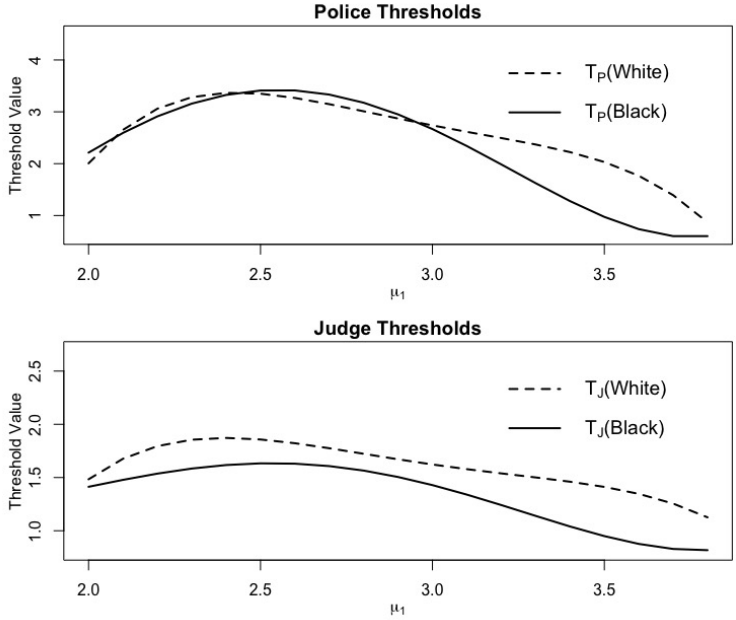
Notes: This figure plots the estimated density of the residualised judge leniency measure constructed using the IV subsample (5656 observations) described in Section 3.2. The residualised judge leniency measure is calculated as the leave-one-out average conviction rate for each judge after accounting for the type of trial and courthouse×year fixed-effects.

Figure B.2: First Stage Monotonicity Check



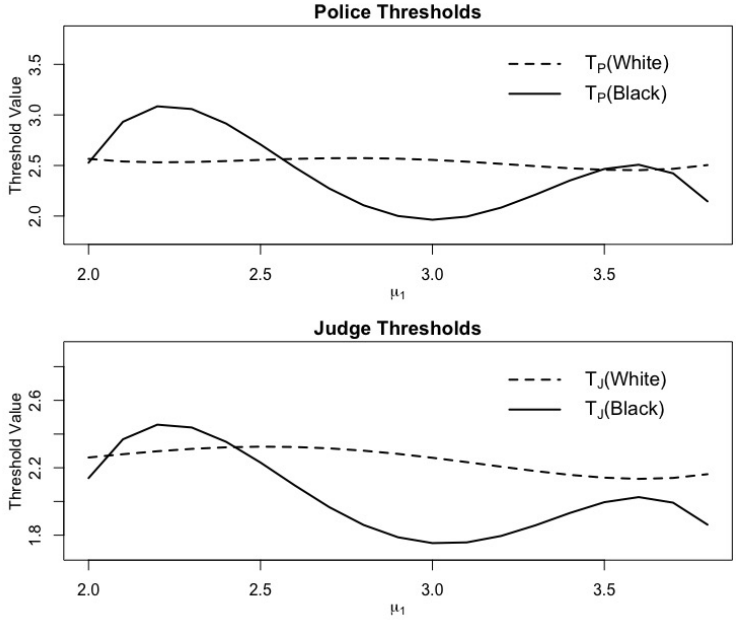
Notes: This figure plots a local-linear-logistic regression of the residualised judge leniency measure on conviction using the IV subsample (5656 observations) described in Section 3.2. The residualised judge leniency measure is calculated as the leave-one-out average conviction rate for each judge after accounting for the type of trial and courthouse×year fixed-effects.

Figure B.3: Calibrated Threshold Values (0.44 Correlation)



Notes: This figure plots the estimated police and judge decision thresholds from the model outlined in Section 5.1 for white and black defendants, respectively. For each race group, the estimated thresholds are obtained by equating the type I and type II error probabilities given in equations (12) and (13) with their respective empirical estimates and solving for the values of $T_P\left(Black\right), T_J\left(Black\right), T_P\left(White\right), T_J\left(White\right)$. In particular, the average empirical estimates of the type I error are 0.25 and 0.17 for blacks and whites, respectively, and 0.18 and 0.16 for the average empirical type II error for blacks and whites, respectively. The correlation between the strength-of-evidence signal seen by the police and the signal seen by the judge is 0.44.

Figure B.4: Calibrated Threshold Values (0.89 Correlation)



This figure plots the estimated police and judge decision thresholds from the model outlined in Section 5.1 for white and black defendants, respectively. For each race group, the estimated thresholds are obtained by equating the type I and type II error probabilities given in equations (12) and (13) with their respective empirical estimates and solving for the values of $T_P\left(Black\right), T_J\left(Black\right), T_P\left(White\right), T_J\left(White\right)$. In particular, the average empirical estimates of the type I error are 0.25 and 0.17 for blacks and whites, respectively, and 0.18 and 0.16 for the average empirical type II error for blacks and whites, respectively. The correlation between the strength-of-evidence signal seen by the police and the signal seen by the judge is 0.89.

Table B.1: OLS Regression

	Dependent variable:			
	Future Criminality			
	(1)	(2)	(3)	(4)
Convicted	0.14***	0.09**	0.05	0.03
	(0.04)	(0.04)	(0.04)	(0.04)
White		0.25***	0.25***	0.24***
		(0.03)	(0.03)	(0.03)
Male		0.28***	0.28***	0.28***
		(0.03)	(0.03)	(0.03)
Previous Criminality		✓	✓	✓
Case Characteristics			✓	✓
ZIP Code Fixed-Effects				✓
Observations	5,656	5,656	5,656	5,656
Adjusted R^2	0.01	0.05	0.06	0.07

Notes: This table reports results from four OLS regressions using the subsample of six circuit courts from Virginia as detailed in Section 3.2. The dependent variable is future criminality calculated using the procedure given in Section 3.1 and is standardised to have unit variance. Conviction is a binary indicator for whether the defendant was convicted. The other regressors are constructed as per the discussion in 3.2. Column (1) gives the simple regression of future criminality on conviction status. Column (2) adds the defendant's race, gender, and seven measures of previous criminality. Column (2) includes whether the crime is an infraction or misdemeanour and the severity of the crime (measured by average fine and average prison sentence, respectively). Column (4) adds ZIP code fixed-effects. Standard errors for the estimated coefficients are reported in parentheses and are clustered at the individual level. * indicates significance at 10%, ** indicates significance at 5%, and *** indicates significance at 1%.

Table B.2: IV Regression (Blacks)

	Dependent variable:			
	Future Criminality			
	(1)	(2)	(3)	(4)
Convicted	0.17	0.09	0.13	0.10
	(0.29)	(0.29)	(0.28)	(0.29)
Male		0.34^{***}	0.34^{***}	0.34^{***}
		(0.04)	(0.04)	(0.04)
Previous Criminality		✓	✓	✓
Case Characteristics			✓	✓
ZIP Code Fixed-Effects				✓
Observations	3,669	3,669	3,669	3,669

Notes: This table reports results from four IV regressions using the subsample of six circuit courts from Virginia as detailed in Section 3.2 for black defendants. The dependent variable is future criminality calculated using the procedure given in Section 3.1 and is standardised to have unit variance. Conviction is a binary indicator for whether the defendant was convicted, it is instrumented by judge leniency. Judge Leniency is the residualised leave-one-out average conviction rate of the judge after controlling for the type of trial and courthouse×year fixed-effects. The other regressors are constructed as per the discussion in 3.2. Column (1) gives the simple IV regression of future criminality on conviction status. Column (2) adds the defendant's race, gender, and seven measures of previous criminality. Column (2) includes whether the crime is an infraction or misdemeanour and the severity of the crime (measured by average fine and average prison sentence, respectively). Column (4) adds ZIP code fixed-effects. Standard errors for the estimated coefficients are reported in parentheses and are clustered at the individual level. * indicates significance at 10%, ** indicates significance at 5%, and *** indicates significance at 1%.

Table B.3: IV Regression (Whites)

	Dependent variable:			
	Future Criminality			
	(1)	(2)	(3)	(4)
Convicted	0.01	-0.01	0.02	-0.04
	(0.35)	(0.34)	(0.37)	(0.40)
Male		0.20**	0.18**	0.16**
		(0.06)	(0.06)	(0.06)
Previous Criminality		✓	✓	✓
Case Characteristics			✓	✓
ZIP Code Fixed-Effects				✓
Observations	1,987	1,987	1,987	1,987

Notes: This table reports results from four IV regressions using the subsample of six circuit courts from Virginia as detailed in Section 3.2 for white defendants. The dependent variable is future criminality calculated using the procedure given in Section 3.1 and is standardised to have unit variance. Conviction is a binary indicator for whether the defendant was convicted, it is instrumented by judge leniency. Judge Leniency is the residualised leave-one-out average conviction rate of the judge after controlling for the type of trial and courthouse×year fixed-effects. The other regressors are constructed as per the discussion in 3.2. Column (1) gives the simple IV regression of future criminality on conviction status. Column (2) adds the defendant's race, gender, and seven measures of previous criminality. Column (2) includes whether the crime is an infraction or misdemeanour and the severity of the crime (measured by average fine and average prison sentence, respectively). Column (4) adds ZIP code fixed-effects. Standard errors for the estimated coefficients are reported in parentheses and are clustered at the individual level. * indicates significance at 10%, ** indicates significance at 5%, and *** indicates significance at 1%.