Using a Special Regressor to Identify Type I and Type II Error Probabilities, With an Illustrative Application to Miscarriages of Justice

Using a Special Regressor to Identify Type I and Type II Error Probabilities, with an Illustrative Application to Miscarriages of Justice

Shin Kanaya*
Department of Economics, University of Essex
Kyoto Institute of Economic Research, Kyoto University
and
Luke Taylor†
Department of Economics and Business Economics, Aarhus University

November 12, 2025

Abstract

We nonparametrically identify type I and type II error rates when mistakes are unobserved. Our strategy reframes the setting as a misclassified binary choice model, using a special regressor to identify the misclassification probabilities. Identification requires an exclusion restriction and a large support condition. When the exclusion restriction is in doubt, we show how our estimands can be interpreted as bounds under a weaker monotonicity condition. Bounds are also provided when the large support condition is violated. We apply our method to estimate miscarriages of justice in Virginia. Among defendants who proceeded to trial, the estimated probability of wrongful acquittal ranges from 18% to 42%, depending on offense type, race, and gender. Our method also produces estimates of wrongful conviction probabilities; however, these are very high, potentially reflecting the restriction of our sample to trial defendants and/or violations of key assumptions.

Keywords: Binary choice, nonparametric identification, criminal justice, misclassification

*Address: Department of Economics, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, United Kingdom. Email: shin.kanaya@essex.ac.uk.

†Address: Department of Economics and Business Economics, Aarhus University, Fuglesangs Allé 4, Aarhus V 8210, Denmark. Email: lntaylor@econ.au.dk.

1 Introduction

There is potential for error — either type I or type II error — whenever a binary decision is made with incomplete information. In this paper, we develop nonparametric identification results that admit simple, intuitive estimators of these error rates when the truth is unobserved. Our approach recasts the problem as a misclassified binary choice model, where the observed outcome is a noisy indicator of the true but unobserved state.

Identification requires a continuous 'special regressor' (Lewbel 1998) that affects the true outcome but is mean-independent of the misclassification process conditional on observables and the true outcome. For settings where this assumption may not hold, we show how the estimands can be interpreted as upper/lower bounds under a weaker condition that allows the misclassification rates to monotonically depend on the special regressor. In addition, the special regressor must satisfy a large support condition akin to that in (Lewbel 2000*a*), meaning that the regressor takes values sufficiently large and small relative to the regression error. To improve the applicability of our estimator, we propose a check of this condition, discuss how the estimates can be interpreted as bounds if the assumption is not met, and develop an alternative identification method based on a certain symmetry of the regression error if only one tail of the special regressor satisfies the large support requirement.

We illustrate our procedure by estimating the probability of a miscarriage of justice using court data from Virginia. Criminal courts have existed for over a thousand years, yet no reliable method exists to measure their performance. Concerns regarding miscarriages of justice date back to antiquity (Zalman 2017), and public interest has grown through media chronicling the (in)famous trials of O.J. Simpson, Steven Avery, and the 'Central Park Five'. However, our method is not limited to courts; it applies to all settings where choices are made with incomplete information and mistakes are unobserved. Examples include hiring decisions, university admissions, promotions, and credit approvals. In such contexts,

downstream outcomes (e.g., grades, earnings, repayment) may serve as special regressors to uncover average error rates.

In our empirical setting, we use a measure of future criminality — the average yearly dollar amount of criminal fines — as the special regressor. The exclusion restriction requires that future criminality is unrelated to the probability of conviction conditional on whether a defendant is guilty. This would be violated if there is an unobservable that is both related to whether they are convicted and to whether they commit crimes in the future, such as the defendant's demeanor at trial. Another potential violation is whether the judge's decision affects future criminality. To address this, we restrict the sample to misdemeanors (where punishments are mild) and to defendants with prior convictions, for whom Agan et al. (2023) find no criminogenic effects. While the conditional mean independence assumption is inherently untestable, we use the conviction tendency of quasi-randomly assigned judges as an IV to show that the effect of conviction on future criminality is small and statistically insignificant.

The large support assumption has two parts: (1) to estimate the false conviction rate, the probability of guilt should approach zero as future criminality approaches its minimum value; and (2) to estimate the false acquittal rate, the probability of guilt should approach one as future criminality approaches its maximum value. The first condition is somewhat implausible *ex ante*, and our diagnostic check confirms it is not satisfied. However, imposing additional assumptions on a certain symmetry of the regression error, we can regain identification and estimate wrongful conviction rates ranging from 40% to 70%, depending on the defendant's offense type, race, and gender. Nevertheless, we caution against placing much weight on these estimates since the symmetry assumption cannot be verified. The second condition of the large support assumption appears satisfied in our check, allowing estimates of wrongful acquittal rates ranging from 18% to 42%.

1.1 Previous Literature

This paper contributes to three literatures. First, we add to the literature on misclassified binary choice models. The first parametric identification results for this model are given by Hausman et al. (1998). Lewbel (2000*a*) extends this to a semiparametric model and uses a special regressor for identification. Although we focus on identifying the misclassification rates — unlike Lewbel (2000*a*) who considers the regression parameters — once those rates are known, the regression parameters can subsequently be identified. Similarly to Lewbel (2000*a*), we also use a special regressor; however, we provide weaker conditions for identification and propose a simpler estimator. Indeed, Lewbel (2000*a*) 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, we use a single nonparametric regression and do not estimate derivatives. Furthermore, we propose a method to check the crucial large support condition that has long been a thorn in the side of special regressor methods. Examples of other papers that apply special regressor methods include Heckman & Navarro (2007) in a dynamic choice model, Berry & Haile (2014) to estimate demand functions, and Lewbel & Tang (2015) and Khan & Nekipelov (2018) in game-theoretic models.

Second, we contribute to estimating the prevalence of miscarriages of justice. Existing work has focused almost exclusively on the probability of wrongful conviction, ignoring wrongful acquittal, and relies on exoneration data (e.g., Gross et al. 2014, Bjerk & Helland 2020). However, exoneration is not equivalent to innocence.

To our best knowledge, Spencer (2007) is the only work (besides ours) that does not rely on exoneration data and estimates the probability of wrongful acquittal. He uses judge–jury agreement rates, but assumes judges and juries make mistakes at the same rate and that their decisions are independent — assumptions that seem implausible. For example, both

judge and jury are more likely to err in complex cases. In contrast, we require data from only one decision-maker.

We also add to the literature on bias in decision-making. In particular, Arnold et al. (2022, ADH hereafter) propose a definition of racial discrimination in bail decisions that is directly linked to our estimand (see Section 2.1). ADH exploit information on judge error rates estimated via extrapolation and an identification-at-infinity assumption regarding a hypothetical 'supremely lenient' judge. We also employ a form of identification-at-infinity using an analogous 'supremely guilty/innocent defendant'. A key distinction is that ADH consider a partially-observed true outcome (pre-trial misconduct) while our true outcome (factual guilt) is unobservable, necessitating a different identification strategy.

2 Identification

2.1 Baseline Identification

Our objects of interest are the type I and type II error probabilities defined as

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

where Y and Y^* are binary-valued variables. Y^* denotes the true unobservable outcome, Y is an observed but (potentially) misclassified version of Y^*, and X is a vector of observable covariates. In our empirical setting, 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 the case and 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.

In comparison to ADH, their measures of discrimination in Equations 1 and 2 (p. 3000) link directly to our $(\alpha_1(x), \alpha_2(x))$ (see also Arnold et al. 2021). Specifically, they define discrimi-

nation as the sum of $E[D|D^*=0, R=w] - E[D|D^*=0, R=b]$ and $E[D|D^*=1, R=w] - E[D|D^*=1, R=b]$, where $R \in (b, w)$ denotes the race (black or white) of the defendant, D is the decision of the judge to release the defendant on bail, and D^* indicates whether there was pre-trial misconduct. The judge aims to match D to $(1 - D^*)$; in our case, the judge aims to match Y with Y^*. Crucially, their D^* is partially observed by the researcher, while our Y^* is unobserved. We also condition on additional regressors X (in line with, e.g., Canay et al. 2022), unlike ADH. This conditioning has important implications for how the discrimination measure is interpreted; however, we do not focus on this nor claim our results imply discrimination of any particular form.

For identification, we assume a special regressor, V, that satisfies the following condition.

Assumption 1 (Exclusion Restriction) *There exists a scalar-valued, continuously distributed variable V that satisfies*

$$E[Y|Y^*, X, V] = E[Y|Y^*, X] \text{ almost surely.} \tag{1}$$

In this setup, $(Y - Y^*)$ is the error, and Assumption 1 can alternatively be written as

$$E[Y - Y^*|Y^*, X, V] = E[Y - Y^*|Y^*, X] \text{ almost surely.} \tag{2}$$

In our empirical setting, the special regressor is future criminality of the defendant, measured as the future average yearly dollar amount of criminal fines, so we require that, on average, the error depends on case and defendant characteristics and their factual guilt but not their future criminality. In Section 6, we provide a thorough investigation into the validity of Assumption 1 for future criminality.[1]

[1]Assumption 1 is not intended to impose a behavioral structure on the judges who decide Y. However, if it were interpreted as a threshold rule as in Canay et al. (2022) and Hull (2021), we could have: $s(X, V, U_1) \geq \tau(X, V, U_2)$ holds iff $Y = 1$, where $s(\cdot)$ is the judge's subjective belief of guilt and $\tau(\cdot)$ the conviction threshold, with (U_1, U_2) unobservable to the researcher. In this framework, it is plausible that V

Under this assumption, we can write

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

Note that the objects of interest, $\alpha_1(x)$ and $\alpha_2(x)$, are independent of v. This expression forms the basis of our identification argument; however, two further assumptions are required. For clarity of exposition, we assume that the support of $V|X=x$ is a bounded and closed interval $[l_V^x, r_V^x]$ for each $x \in supp\,(X)$, where $supp\,(X)$ denotes the support of X.

Assumption 2a (Lower Tail Condition) *For each $x \in supp\,(X)$,*

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

Assumption 2b (Upper Tail Condition) *For each $x \in supp\,(X)$,*

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

Assumption 2a requires that defendants with no future criminality are factually innocent—a condition that is shown to be false in our empirical application. While Assumption 2b requires that defendants with extremely high future criminality are factually guilty—our empirical check appears to validate this.[2] Equivalently, we can write this in terms

neither directly enters the subjective belief nor the threshold as it is a future variable. Further, if $(U_1, U_2) \perp V|X, Y^*$, Assumption 1 holds.

[2]With minor modifications, all results hold with an unbounded or (semi-) open interval, including $(-\infty, \infty)$. Indeed, Assumptions 2a and 2b are written using limit notation such that the conditions hold for unbounded or open support 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 $[l_V^x, r_V^x]$ (and the conditional probability is continuous at the boundary points of the support).

of the support of $V|X = x$ as follows. Let $Y^* = \mathcal{I}(m(X,V) \geq U)$, where $\mathcal{I}(\cdot)$ is the indicator function, $m(\cdot,\cdot)$ an unknown function, and U an unobservable random variable with $U \perp V|X$. Then we require the support of $m(X,V)|X = x$ to contain the support of $U|X = x$; thus, we refer to these assumptions jointly as a large support condition. They can also be viewed similarly to the 'supremely lenient judge' assumption of ADH.

From Equation (3), Assumption 2a identifies $\alpha_1(x)$, and Assumption 2b identifies $\alpha_2(x)$.

Theorem 1 *Under Assumptions 1 and 2a*

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

Theorem 2 *Under Assumptions 1 and 2b*

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

When the support of $V|X = x$ is $(-\infty, \infty)$, this identification is known as 'identification-at-infinity'. However, in our setting, V is the fine amount, so is bounded from below by 0 and has a finite upper bound since the maximum fine for any crime is \$250; thus, this problem concerns boundary estimation, and is more akin to regression discontinuity design. We note that the distinction between infinite and bounded support cases is not substantive for our identification result since one can always redefine V as $\tilde{V} = \Phi(V)$ using the normal distribution function Φ, for example, and this $\tilde{V}$ has bounded support.[3]

Under the assumption that the support of $V|X = x$ is a bounded interval, $[l_V^x, r_V^x]$, we can think of two cases for the endpoint r_V^x (analogous arguments hold for l_V^x): the (conditional) probability density of V satisfies (i) $f_{V|X}(r_V^x|x) > 0$; or (ii) $f_{V|X}(r_V^x|x) = 0$ (the support of $V|X = x$ is defined as the smallest closed set A on which $P[V \in A|X = x] = 1$, where the closedness is defined on $\mathbb{R}$). For Case (i), even if $P[Y^* = 1|(X,V) = (x,v)] < 1$ for any $v <$

[3]The discussion on this point and that in the following paragraph were made possible by insightful suggestions by the Editor, Associate Editor, and a referee.

r_V^x and it first attains value 1 at $v = r_V^x$, $\alpha_2(x)$ can be identified and estimated in a standard manner (e.g., using the local linear logit estimator, which achieves the same convergence rate at boundary and interior points). In contrast, for Case (ii) with $f_{V|X}(r_V^x|x) = 0$, the conditional probability $P[Y^* = 1|(X, V) = (x, r_V^x)]$ is not well defined. Even in this case, identification is still possible provided that $\lim_{v \to r_V^x} P[Y^* = 1|(X, V) = (x, v)]$ exists and equals 1. However, this corresponds to an irregular or thin-set identification problem as discussed in Khan & Tamer (2010) and Andrews & Schafgans (1998) (where the latter paper's setting corresponds to $r_V^x = \infty$ in our context, so that the conditional probability is not well defined at $r_V^x = \infty$ and the density value tends to zero as $v \to r_V^x = \infty$, such that identification is based on 'extrapolation' for the endpoint r_V^x). We do not pursue this irregular identification case here, but conjecture that upon construction of a suitable estimator, its convergence rate depends on the tail behavior of $f_{V|X}(v|x)$ and related quantities (as $v \to r_V^x$), as observed in the aforementioned papers.[4]

If one believes that either or both of the tail conditions are only approximately satisfied, the estimators given by Theorems 1 and 2 are upper bounds on the true error rates. Indeed, suppose $\lim_{v \to l_V^x} P\left[Y^* = 1 \middle| (X, V) = (x, v)\right] = \epsilon$ for some small $\epsilon > 0$, then

$$\lim_{v \to l_V^x} P\left[Y = 1 \middle| (X, V) = (x, v)\right] \quad = \quad \alpha_1(x) + \left[1 - \alpha_1(x) - \alpha_2(x)\right] \epsilon \geq \alpha_1(x), \tag{8}$$

since $[1 - \alpha_1(x) - \alpha_2(x)] > 0$. Similarly, if $\lim_{v \to r_V^x} P\left[Y^* = 1 \middle| (X, V) = (x, v)\right] = 1 - \tilde{\epsilon}$ for

[4] Goh (2018) considers a CDF transformation of a conditioning variable, which corresponds to replacing the conditioning variable, V, with $F_{V|X}(V|x)$ in our case, and shows that his estimator based on that transformation does not depend directly on the tail behavior of the density of the original variable (corresponding to our $f_{V|X}(v|x)$). We conjecture that a comparable result may hold in our context.

some small $\tilde{\epsilon} > 0$, then

$$\begin{aligned} 1 - \lim_{v \to r_V^x} P\left[Y^* = 1 | \left(X, V\right) = (x, v)\right] &= 1 - \alpha_1(x) - \left[1 - \alpha_1(x) - \alpha_2(x)\right](1 - \tilde{\epsilon}) \\ &= \alpha_2(x) + \left[1 - \alpha_1(x) - \alpha_2(x)\right]\tilde{\epsilon} \\ &\geq \alpha_2(x). \end{aligned} \tag{9}$$

2.2 Relaxing the Exclusion Restriction

In Section 6, we provide empirical support for the validity of the exclusion restriction. However, these are only checks, not proofs, and cautious readers may still be concerned about the validity of this assumption. As such, we explore what can be learned from our results if we relax the exclusion restriction. In particular, suppose:

Assumption 1′ (Alternative Exclusion Restriction) *There exists a scalar-valued, continuously distributed variable V that satisfies*

$$E\left[Y | Y^*, X, V = v\right] \geq E\left[Y | Y^*, X, V = v'\right] \text{ almost surely for } v > v'. \tag{10}$$

This monotonicity condition substantially relaxes the exclusion restriction. It says that V can affect the conditional mean of Y but may do so only in a monotonically increasing way. Intuitively, judges should be more likely to convict a defendant with higher future criminality than one with lower (of course, this need only hold on average for each (Y^*, X)).

Replacing our exclusion restriction with this monotonicity condition results in the following adjustment to Equation (3):

$$\begin{aligned} &P\left[Y = 1 | (X, V) = (x, v)\right] \\ &= \alpha_1\left(x, v\right) + \left[1 - \alpha_1\left(x, v\right) - \alpha_2\left(x, v\right)\right] P\left[Y^* = 1 | (X, V) = (x, v)\right], \end{aligned} \tag{11}$$

where

$$\alpha_1(x,v) := P\left[Y=1|Y^*=0, X=x, V=v\right], \tag{12}$$

$$\alpha_2(x,v) := P\left[Y=0|Y^*=1, X=x, V=v\right]. \tag{13}$$

Note that because Assumption 1′ must hold for both values of Y^* (and all values of X), using the previous definitions of $\alpha_1(x,v)$ and $\alpha_2(x,v)$, we can write Assumption 1′ equivalently as

$$\alpha_1(X,v) \geq \alpha_1(X,v') \text{ almost surely for } v > v', \tag{14}$$

$$\alpha_2(X,v) \leq \alpha_2(X,v') \text{ almost surely for } v > v'. \tag{15}$$

Thus, using the same method of identification as in Section 2.1, under Assumption 1′ and Assumptions 2a and 2b, respectively, we have

$$\lim_{v \to l_V^x} P\left[Y=1|\left(X,V\right)=(x,v)\right] = \alpha_1(x,l_V^x) \leq \alpha_1(x,\tilde{v}), \text{ for all } \tilde{v}, \tag{16}$$

$$1 - \lim_{v \to r_V^x} P\left[Y=1|\left(X,V\right)=(x,v)\right] = \alpha_2(x,r_V^x) \geq \alpha_2(x,\tilde{v}), \text{ for all } \tilde{v}. \tag{17}$$

We can interpret $\alpha_1(x,l_V^x)$ as the 'best case scenario', in that the probability of convicting an innocent defendant with characteristics x is at least $\alpha_1(x,l_V^x)$, while the probability of acquitting a guilty defendant is no larger than $\alpha_2(x,r_V^x)$.

2.3 Checking the Tail Conditions

The tail conditions (Assumptions 2a and 2b) are critical for identification. Thus, it is important to provide a method to empirically check their validity. To this end, we introduce an additional assumption:

Assumption 3 (Single-Index Structure)

$$Y^* = \mathcal{I}\left(V + h(X) \geq U\right), \tag{18}$$

where $h(\cdot)$ 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, $F_{U|X}(\cdot|x)$, *has a corresponding density* $f_{U|X}(\cdot|x)$.

The conditional independence condition $U \perp V|X$ in Assumption 3 resembles that of Lewbel (2000*b*, Assumption A2). This, together with the single-index structure of Y^* in Equation (18), leads to the following expression for the 'conditional predictive probability' (CPP):

$$P[Y^* = 1|(X, V) = (x, v)] = F_{U|X}(v + h(x)|x), \tag{19}$$

where we note that $F_{U|X}(\cdot|x)$ is also defined on the region outside the support of $U|X = x$ as $F_{U|X}(s|x) = 0$ $(= 1)$ for any s that is smaller (larger) than any points in the support of $U|X = x$ (with its corresponding density value $f_{U|X}(s|x) = 0$), so that the left-hand side of Equation (19) is well defined even when $v + h(x)$ is not in the support of $U|X = x$.

While Assumption 3 may look restrictive, any CPP that is monotone in v can be represented by the model in Assumption 3 under mild regularity conditions (given as Theorem 4 in the appendix; cf. Section 2 of Magnac & Maurin 2007).

Note that this single-index model is not structural, i.e. it does not attempt to explain a defendant's behavior. It is merely a tool for the researcher to predict such behavior retrospectively. This stands in contrast to previous work that uses similar single-index specifications and conditional independence assumptions to create structural models (e.g., Berry & Haile 2014), where careful consideration must be given to the underlying behavioral mechanisms that could generate such a model.

When Assumption 3 is imposed, the two tail conditions (Assumptions 2a and 2b) together imply the support of $V|X = x$ contains that of $(U - h(X))|X = x$, i.e. $[l_U^x - h(x), r_U^x - h(x)] \subseteq [l_V^x, r_V^x]$, where l_U^x and r_U^x are the left and right endpoints of the support of $U|X = x$.

To construct our check, first note that $F_{U|X}(l_V^x + h(x)|x) = 0$ is a necessary and sufficient condition for Equation (4) to hold; equally $F_{U|X}(r_V^x + h(x)|x) = 1$ is a necessary and

sufficient condition for Equation (5) to hold. Under Assumptions 1 and 3, the partial derivative of Equation (3) with respect to v is

$$\frac{\partial}{\partial v} E\left[Y|V=v, X=x\right] = \left[1-\alpha_1(x)-\alpha_2(x)\right] f_{U|X}(v+h(x)|x). \tag{20}$$

The left-hand side of Equation (20) can be directly estimated from the data. Furthermore, for a given x, the right-hand side is a constant multiple of $f_{U|X}(v+h(x)|x)$.

If the derivative is zero in a neighborhood of the upper limit of $V|X=x$ (i.e., for any $v \in [r_V^x - \varepsilon, r_V^x]$ with some $\varepsilon > 0$), this suggests $F_{U|X}(r_V^x + h(x)|x) = 1$ under the condition that $f_{U|X}(\cdot|x)$ is strictly positive in the interior of $[l_U^x, r_U^x]$. Equally, a zero derivative in a neighborhood of the lower limit of $V|X=x$ indicates $F_{U|X}(l_V^x + h(x)|x) = 0$ under the same condition. In our empirical application, Section 4 checks whether there is a neighborhood of each endpoint where the derivative estimates are zero, which can be interpreted as examining a sufficient condition for each tail condition for each point of interest x.

To see why our check is only sufficient, note that $f_{U|X}(r_V^x + h(x)|x) = 0$ is neither sufficient nor necessary for $F_{U|X}(r_V^x + h(x)|x) = 1$ (analogously for l_V^x), since knowing the derivative at a point does not determine the level of the function. For simplicity, suppose $h(x) = 0$, and consider two cases:

(a) If the upper limit of $V|X=x$, r_V^x, is strictly smaller than that of $U|X=x$, we can still have $f_{U|X}(r_V^x|x) = 0$ but $F_{U|X}(r_V^x|x) < 1$.

(b) If $V|X=x$ and $U|X=x$ are uniformly distributed on $[0,1]$, so their supports coincide, $f_{U|X}(r_V^x|x) = 1$ and $F_{U|X}(r_V^x|x) = 1$.

Cases like (a) are ruled out by the strict positivity of $f_{U|X}(\cdot|x)$, while cases like (b), though yielding $F_{U|X}(r_V^x|x) = 1$, do not pass our check. Together, these examples clarify why strict positivity is essential and why our check ensures only sufficiency.[5]

[5]These arguments follow from the Editor's insightful comments.

So-called falsification tests, often used in empirical work, typically examine a testable implication that is necessary but not sufficient for the validity of an underlying assumption.[6] Our zero-derivative check, in contrast, examines a sufficient (though not necessary) condition for the tail condition (under the untestable assumption of the strict positivity of $f_{U|X}(\cdot|x)$). Thus, unlike typical falsification tests, its failure does not immediately imply a violation of that condition (as in (b) above).

2.4 Relaxing the Tail Conditions

Here, we consider an alternative identification scheme for cases in which Assumptions 2a and 2b do not hold simultaneously, as in our application (see Section 4). Without loss of generality, we proceed without Assumption 2a and impose the following assumption.

Assumption 4 (Mode-Median Coincidence/Limited Predictability) *There exists a unique global 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 4 is satisfied by commonly-used parametric distributions; its simplest sufficient condition is that $U|X = x$ is symmetric and unimodal, as in the case of the Gaussian or logistic distributions; however, it does not exclude non-symmetric distributions.[7]

We interpret Assumption 4 as a limited predictability condition in the following way. From

[6] For example, in randomized controlled trials, balance on observable characteristics between treated and control groups is a necessary implication of random assignment but not sufficient to ensure that randomization was correctly implemented. This example and the explanatory note here are motivated by the Associate Editor's constructive suggestion.

[7] Note that the maximum point m_U^x need not necessarily be the mode of $U|X = x$, i.e. the true mode may exist outside of $[l_V^x + h(x), r_U^x]$. We simply require that the unique maximum point inside $[l_V^x + h(x), r_U^x]$ equals the median; nonetheless, we maintain the mode interpretation for ease of understanding.

the form of the CPP in Equation (19), the median of $U|X = x$ occurs where the probability of the defendant being guilty is 0.5. Since we require Mode $[U|X = x]$ = Median $[U|X = x]$, there must be 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. Importantly, note that this is not required to hold in the population as a whole. We only require, for a given point of interest x, some value of V for which this holds.

We can also interpret Assumption 4 as a location normalization. Manski (1988) imposes a location normalization through a conditional median restriction to identify $h(x) = x'\beta$. In our context, his assumption corresponds to Median $[U|X] = 0$. While Mode $[U|X] = 0$ can play the same role, it is important to note that Manski (1988) considers an observable binary outcome. If Y^* were observable in our setting, either Mode $[U|X] = 0$ or Median $[U|X] = 0$ could be used to identify $h(x)$.[8,9] So, Assumption 4 is stronger than necessary when Y^* is observable. Theorem 4 in the appendix gives a representation result for the CPP and clarifies that indeed Mode $[U|X]$ = Median $[U|X]$ imposes more structure on the CPP in Equation (19) than either Mode $[U|X] = 0$ or Median $[U|X] = 0$. 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 restores the identification of $\alpha_1(x)$ when Assumption 2a does not hold. Recall that

$$P\left[Y = 1|\left(X, V\right) = (x, v)\right] = \alpha_1(x) + \left[1 - \alpha_1(x) - \alpha_2(x)\right] F_{U|X}\left(v + h(x)|x\right). \tag{21}$$

[8]It is not necessary that these conditional quantities equal 0; any known number $c_x \in \mathbb{R}$ for each x normalization can be used instead (see Theorem 4 in the appendix).

[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 maximizer v in the support of V for each x (see Theorem 4 in the appendix).

Taking the partial derivative with respect to v gives

$$\frac{\partial}{\partial v} P\left[Y=1|\left(X,V\right)=(x,v)\right]=\left[1-\alpha_1(x)-\alpha_2(x)\right] f_{U|X}\left(v+h(x)|x\right). \tag{22}$$

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

$$\begin{aligned}\bar{v}(x) &:= \text{argmax}_{v\in\left[l_V^x, r_V^x\right]} \frac{\partial}{\partial v} P\left[Y=1|\left(X,V\right)=(x,v)\right] \\ &= \text{argmax}_{v\in\left[l_V^x, r_V^x\right]} f_{U|X}\left(v+h(x)|x\right). \end{aligned} \tag{23}$$

It is straightforward to estimate $\bar{v}(x)$ since $\frac{\partial}{\partial v} P\left[Y=1|\left(X,V\right)=(x,v)\right]$ is identified directly from the data. The restriction $\alpha_1(x)+\alpha_2(x)<1$ is what Hausman et al. (1998) call the monotonicity condition and is standard for misclassified binary variables. 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 — an odd situation.

Here $[\bar{v}(x)+h(x)]$ is the unique maximizer of $f_{U|X}\left(v+h(x)|x\right)$, hence, it is the modal point of $U|X$. Moreover, under Assumption 4, $[\bar{v}(x)+h(x)]$ is the median of $U|X=x$, so $F_{U|X}\left(\bar{v}(x)+h(x)|x\right)=1/2$. Armed with this, we can write

$$\begin{aligned} P\left[Y=1|\left(X,V\right)=(x,\bar{v}(x))\right] &= \alpha_1(x)+\left[1-\alpha_1(x)-\alpha_2(x)\right] F_{U|X}\left(\bar{v}(x)+h(x)|x\right) \\ &= \alpha_1(x)+\left[1-\alpha_1(x)-\alpha_2(x)\right]/2 \\ &= \left[1+\alpha_1(x)-\alpha_2(x)\right]/2. \end{aligned} \tag{24}$$

After obtaining $\alpha_2(x)$ via Theorem 2, we can then obtain $\alpha_1(x)$ since the left-hand-side of Equation (24) is directly identified from the data. We summarize this result in the following theorem.

Theorem 3 *For $\bar{v}(x)$ defined in Equation (23), if $\alpha_1(x)+\alpha_2(x)<1$ for each $x\in supp\left(X\right)$ and Assumptions 1, 3, and 4 hold, then $\alpha_1(x)$ and $\alpha_2(x)$ satisfy*

$$\alpha_1(x) = 2P\left[Y=1|\left(X,V\right)=(x,\bar{v}(x))\right]+\alpha_2(x)-1. \tag{25}$$

3 Institutional Setting and Data

3.1 Institutional Setting

This study uses data from Virginia's 32 General District Courts (GDCs), which preside over all infractions and misdemeanors.[10] Infractions carry no prison sentence and a maximum fine of $250; misdemeanors can be punished by up to one year in prison. GDCs conduct bench trials only — no juries are involved. Our dataset contains only individuals who have been arrested and charged with a crime. Once charged, defendants face a preliminary hearing, where a judge (not necessarily the trial judge) may dismiss the case. If the case proceeds, the defendant may plead guilty and receive a lesser sentence. Otherwise, the trial ensues and the judge determines both guilt and the penalty.

GDCs also hear civil cases, where conviction requires only 50% certainty. In contrast, criminal cases require proof beyond a reasonable doubt. This 'doubt' is the type I error with which we are concerned: the probability of convicting an innocent defendant. Accordingly, our analysis is restricted to criminal cases.

3.2 The Special Regressor

Given its critical role in our analysis, we first discuss the choice of the special regressor: future criminality. There are several viable measures: number of arrests, convictions, dollar amount of fines, or days incarcerated. Recall that we want a variable highly correlated with true guilt so that it satisfies the large support assumption. Using our heuristic check of the large support condition, we select the dollar amount of fines received in the future, which best satisfies the upper tail condition (results are given in Section 4). Intuitively, knowledge that a defendant will have significant involvement with the justice system provides information

[10] Humphries et al. (2023) use a similar dataset from Virginia's Circuit Courts.

about their likely guilt of the current crime, i.e., the joint event 'factually innocent in the current case and repeatedly convicted thereafter' seems unlikely.

Although multiple versions of V were considered, our selection is assumption-driven, not result-driven. That is, the chosen special regressor was not selected based on yielding 'favorable' estimates of type I or type II errors, but rather on its empirical support for the identification assumptions. This differs from data snooping (or p-hacking-like procedures): we aim to validate assumptions, not optimize results. With a large sample size, distortions to estimation and inference introduced by variable selection may be mitigated, or at least are expected to diminish asymptotically.[11] Nonetheless, our inference results (e.g., the reported standard errors of our estimates) should be interpreted with caution as they do not reflect the distributional impact of this step.

Since future criminality occurs after the trial, it cannot influence the probability of a miscarriage of justice. Nonetheless, conditional on true guilt and observed controls, future criminality must be unrelated to unobservables that affect the probability of misclassification. We argue that controlling for previous criminality allays much of this concern. Consider an overall criminality measure composed of previous and future criminality. This measure is likely correlated with unobserved characteristics, such as visible gang tattoos, that also affect judicial decisions. These unobservables can be split into time-invariant and time-varying components. Since we control for previous criminality, we implicitly control for these time-invariant unobservables, leaving only the smaller set of time-varying unobservables

[11] Indeed, if we can formalize a statistical checking procedure that is consistent and has asymptotic size zero (i.e. it eliminates any variable that does not satisfy the identification assumptions and detects one that does when it exists, with probability approaching one), then the estimation results based on the variable that passes the check are not systematically distorted, in that the estimators are consistent and inference is asymptotically valid. While it may be possible to develop such a formal procedure, as in Andrews (2000), who proposes a pretest based on the law of the iterated logarithm, we leave this for future research.

that pose a problem. Moreover, any such unobservable would need to affect the judge's decision conditional on true guilt, and thus represents judicial bias — something unlikely to be predictive of future criminality. If one still believes such an unobservable exists, the interpretation of the misclassification rates becomes specific to those with extreme future criminality. For example, if all those with high future criminality are also those with visible gang tattoos, then the false acquittal rate becomes specific to those with such tattoos.

It is also possible that the judge's decision causally affects future criminality conditional on true guilt and the regressors, violating the mean independence condition. To mitigate this, we restrict the sample to misdemeanors, where punishments are lower, so reducing the effect of a conviction. The mean prison sentence for those convicted in our sample is 49 days, and the median is no prison time at all. Although this removes some of the more interesting offenses, 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.

In addition, we also restrict attention to those with at least one prior conviction or arrest. Agan et al. (2023) find that while the effect of prosecution on future criminal behavior is significant and positive for misdemeanors, they do not find an effect for those with a prior criminal history.

Notwithstanding these data restrictions, in Section 6, we formally test whether the judge's decision Y affects future criminality V using the conviction tendency of the randomly-assigned judge as an IV for the likely endogenous Y. Note that this is not a perfect test of our assumption, as we cannot control for the true guilt of the defendant.

Finally, if there is still concern regarding the conditional mean assumption, Section 2.2 provides bounds on the true probabilities under monotonic violations of this assumption.

3.3 Sample and Variable Construction

The data cover all cases with charges filed between 2009-2020 — more than 20 million — where the unit of observation is a single charge. All non-Virginia residents are dropped (22% of observations dropped). All parole violations are excluded, and only misdemeanors for those with prior arrests or convictions are considered (75% dropped). To address dependence across charges, we remove multi-charge trials, keeping only observations where the defendant faces a single charge (28% dropped). We also restrict to the two most common offense types in the dataset: traffic violations and violent crimes (26% dropped).

As our analysis concerns final trial decisions, we exclude all cases dismissed before trial or resolved by guilty plea (31% dropped). In plea cases, the defendant voluntarily accepts guilt, and the judge plays no role in the final verdict. If a defendant pleads guilty, they are always convicted, so a mistake would never be made if the defendant was truly guilty and would always be made if they were innocent.

Since data are only observed until the end of 2020, future criminality is measured as the average yearly dollar amount of fines received per day after the current trial (or after release if convicted), until the end of the sample period. We construct an analogous measure for previous criminality using fines incurred before the current trial. Having calculated these variables, the first and last years of the data are dropped as 'burn-in periods' (13% dropped). The final sample contains 652,746 observations.

To avoid nonparametric fixed-effects estimators, we compute a ZIP code pseudo-fixed-effect using the leave-one-out average conviction rate for each defendant's home ZIP code. There are 1007 ZIP code areas in Virginia with a mean population of 648 and a median of 145.

We close this section with a final remark. Since the sample includes only individuals who were arrested and charged, our estimates are conditional on this selection. That is, we

estimate

$$\alpha_1(x) = P\left[Y = 1 \middle| Y^* = 0, X = x, A = 1\right], \tag{26}$$

$$\alpha_2(x) = P\left[Y = 0 \middle| Y^* = 1, X = x, A = 1\right], \tag{27}$$

where $A = 1$ indicates that the individual was arrested and charged. For brevity, we suppress the dependence on A. However, it is crucial to recognize that our estimates do not generalize to the general population, but rather apply to the subpopulation already filtered by law enforcement and prosecutorial discretion.

3.4 Descriptive Statistics

Table 1 reports descriptive statistics. There is little racial or gender difference in those acquitted versus convicted. The sample is predominantly male, and, consistent with Virginia's demographic composition (approximately 56% non-Hispanic white), is also majority white.

4 Empirical Check of the Tail Conditions

We assess the tail conditions using the approach in Section 2.3. Specifically, we estimate the nonparametric functions using a local linear likelihood estimator with a logistic link function (see Loader 2006), a tricube kernel, and 25% nearest neighbors (results are qualitatively similar for 15% and 35% nearest neighbors). Frölich (2006) shows in simulations that local linear logit outperforms the Nadaraya-Watson, the local linear kernel, the parametric logit, and Klein & Spady (1993) estimators. It also has better boundary bias properties (Fan et al. 1995), which is important since our identification hinges on the behavior of $E[Y|X, V]$ at the boundaries of V.

We report results by crime type (traffic or violent), race, and gender, controlling for neighborhood-level criminality and previous criminality, with results reported at their 25th

Table 1: Descriptive Statistics

	Convicted		Acquitted	
	Mean	St. Dev.	Mean	St. Dev.
Future Criminality	0.14	0.23	0.11	0.20
Previous Criminality	0.23	0.35	0.19	0.28
Neighborhood Effect	0.80	0.05	0.79	0.06
Male	0.67	0.47	0.63	0.48
Black	0.44	0.50	0.43	0.50
Traffic Crime	0.83	0.38	0.65	0.48
Violent Crime	0.17	0.38	0.35	0.48
Observations	468 386		184 360	
Percentage	71.8%		28.2%	

percentile. Using higher percentiles (e.g., medians) shifts our view of the density to the right and we lose the modal point.

Figure 1 plots estimates of $(\partial/\partial v)\, E\left[Y|X=x, V=v\right]$ over the support of V (future criminality), which should not be mistaken for estimates of the density of V. The upper tail condition appears satisfied for all subgroups. And we note that the derivative is non-negative, providing evidence for the single-index structure (Assumption 3). In the appendix, we directly check this assumption by plotting $E[Y|X=x, V=v]$ in Figure 3; monotonicity of this function implies monotonicity of $P[Y^*=1|X=x, V=v]$ by Equation (3). Note also that from Figure 1, the zero derivative value is already attained before reaching the upper boundary point; thus our check here can be interpreted as passing a sufficient condition of Equation (5) (when the strict positivity of the density $f_{U|X}(\cdot|x)$ is assumed, as discussed in

Section 2.3). In the appendix, Figure 4 shows that the density of V is small in the upper tail, dropping to 0.00003 in the worst case (black females charged with a violent crime). However, the large support check shows the upper tail condition is approximately satisfied for future criminality above 1; at this point, the density of V is 0.015 in the worst case. This suggests the (nearly) zero derivative estimates in Figure 1 are computed reliably (i.e., without an absence of data for V) as well as that identification through the upper tail is unlikely to be a case of thin-set identification (cf. Section 2.1).

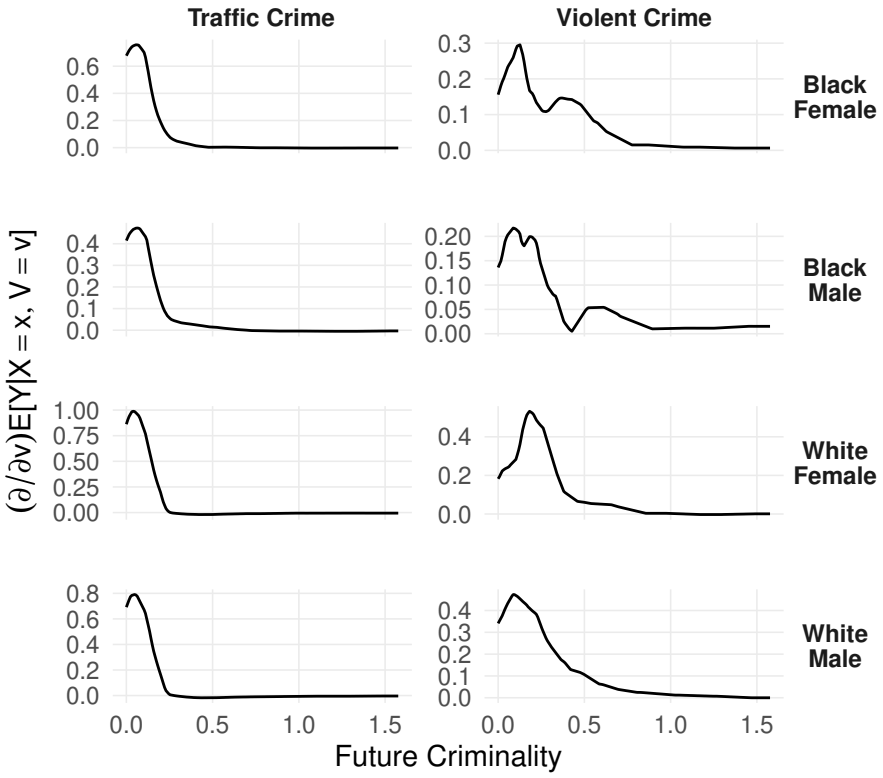

Figure 1: Estimates of $(\partial/\partial v)E\left[Y|X=x,V=v\right]$.

Unfortunately, the lower tail condition is not satisfied, reflecting that we cannot be sure of the innocence of defendants who have no future criminal fines. Nevertheless, it appears the modal point is observed in each case, though the densities for the violent crime cases are more erratic, perhaps reflecting the smaller sample size. If one is willing to assume the median and mode are equal — satisfied if the distributions are symmetric — the

identification scheme of Section 2.4 can be used to recover the probability of incorrect conviction, α_1. However, caution should be applied when arguing for symmetry since so little of the left tail is observed.

5 Results

Table 2 reports the local logit estimates of the probabilities of incorrect conviction (α_1) and incorrect acquittal (α_2) after the final trial, conditional on neighborhood and prior criminality at their 25th percentiles. The estimation procedure and controls follow Section 4, which shows that the large support condition holds for the upper tail but not the lower. However, the figures also indicate the mode is identified, allowing us to use the estimator from Section 2.4 to estimate α_1 without the lower tail condition. These estimates should be interpreted cautiously, as they also rely on a symmetry assumption in the error density that we cannot verify here. Standard errors (in parentheses) are based on a smoothed bootstrap, whose validity for local likelihood estimators is shown in Claeskens & Van Keilegom (2003). As explained in Section 3.3, V is selected based on assumption checks; while the statistical uncertainty from these checks should be accounted for, doing so is difficult and not pursued here. Thus, the reported standard errors should be interpreted with caution. This issue is analogous to the post-model-selection inference problem, whose solutions are known only for specific cases (see, e.g., Bachoc et al. 2020, for simple regression).

Since the estimated probability of wrongful conviction, α_1, hinges on the unverified equality of the error term's median and mode, all α_1 values—especially those for violent offenses, which rely on a much smaller sample—should be viewed with caution. Nevertheless, the rates are striking: they range from 0.40 for white-male violent cases to 0.69 for black-male traffic cases. However, recall that these are conditional on arrest and charge; anyone reaching trial has already faced a stringent evidentiary hurdle, so these estimates will be much higher

Table 2: Type I and II Probability Estimates

	α_1		α_2	
	Traffic	Violent	Traffic	Violent
Black Female	0.570 (0.013)	0.457 (0.037)	0.178 (0.006)	0.418 (0.011)
Black Male	0.696 (0.008)	0.483 (0.018)	0.176 (0.003)	0.417 (0.008)
White Female	0.566 (0.016)	0.543 (0.037)	0.182 (0.004)	0.315 (0.010)
White Male	0.637 (0.011)	0.422 (0.041)	0.184 (0.003)	0.383 (0.008)

than for the general population.

Offense type appears to drive the biggest differences. Defendants charged with traffic offenses are far more likely to be wrongfully convicted (0.56–0.69) and far less likely to be wrongfully acquitted (0.17–0.18) than those tried for violent crimes ($\alpha_1 \approx 0.4 - 0.55$ and $\alpha_2 \approx 0.31 - 0.42$). This pattern is consistent with a higher evidentiary threshold in violent cases.

There are no obvious patterns in the effect of gender or race in these miscarriage of justice probabilities. From a statistical standpoint, the α_2 estimates for traffic offenses are the most trustworthy: they come from the largest subsample, require no median-mode equality assumption, and carry the smallest standard errors. Their uniformity across race and gender likely reflects the objective, often photographic, proof typical of traffic violations.

Our estimates of the probability of wrongful acquittal, $\alpha_2(x)$, are obtained under the baseline exclusion restriction (Assumption 1) and the upper-tail condition (Assumption 2b). When the exclusion restriction is weakened to monotonicity (Assumption 1′), these estimates should instead be interpreted as upper bounds, as shown in Equation (17). For instance, we find that *at most* 17.8% of black female defendants are wrongfully acquitted.

By contrast, our estimates of the probability of wrongful conviction rely on the mode–median coincidence condition (Assumption 4). Under this identification scheme, relaxing the exclusion restriction to monotonicity removes the ability to interpret these estimates as bounds, since the maximiser $\bar{v}(x)$ (defined in the first line in Equation (23)) can no longer be identified and estimated as the modal point.

6 Validity of the Exclusion Restriction

As discussed in Section 3.2, the exclusion restriction (Assumption 1) is violated if miscarriages of justice affect future criminality. Although we cannot directly test this, since it requires knowledge of the defendant's true guilt Y^*, we can test whether observed conviction, Y, affects future criminality. Recent studies suggest that convictions have a criminogenic effect. Both Kamat et al. (2023) and Agan et al. (2023) find that misdemeanor convictions increase the likelihood of reoffending. However, Agan et al. (2023) also find no effect for defendants with a prior criminal record. If there is indeed a criminogenic effect, our conditional mean independence condition is violated but our monotonicity condition in Assumption 1' would be satisfied, consistent with findings in the aforementioned two papers.

To investigate this question in our context, we use the conviction tendency of quasi-randomly assigned judges as an instrument for conviction.

6.1 IV Institutional Setting and Data

We restrict our analysis to Chesterfield County General District Court, where personal correspondence with the court clerk confirms that judges are randomly assigned to cases. While we lack similar confirmation for other courts, restricting attention to a single court to increase case homogeneity is conventional (e.g., Agan et al. 2023). Chesterfield GDC is staffed by five judges at any given time, each elected by the Virginia General Assembly for

six-year terms.

Table 3: Descriptive Statistics (IV Subsample)

	Convicted		Acquitted	
	Mean	St. Dev.	Mean	St. Dev.
Judge Severity	0.82	0.05	0.81	0.04
Future Criminality	0.15	0.22	0.12	0.19
Previous Criminality	0.25	0.31	0.20	0.27
Neighborhood Effect	0.76	0.02	0.76	0.02
Male	0.67	0.47	0.63	0.48
Black	0.52	0.50	0.50	0.50
Traffic Crime	0.85	0.36	0.64	0.48
Violent Crime	0.15	0.36	0.36	0.48
Observations	24 238		5 306	
Percentage	82%		18%	

The sample size for this analysis is 29,544, with 46 judges hearing an average of 642 cases each (median: 551). We calculate severity for each judge in each year via the UJIVE estimator of Kolesár (2013), which is consistent in the face of heterogeneous treatment effects, whereas standard leave-one-out IV is not.[12] Moving from a judge one standard deviation below the mean to a judge one standard above the mean increases conviction probability by 9.8 percentage points, relative to a mean conviction rate of 82% in this subsample. Table 3 provides descriptive statistics, which are broadly similar to those for

[12]Our severity measure is constructed in the same manner as the residualized leave-one-out ADA leniency measure in Agan et al. (2023), following standard practice in the literature. Estimation is carried out using the 'jive' R package from Kyle Butts.

the full sample, suggesting we can generalize our findings.

6.2 IV Analysis

Columns (1) - (3) of Table 4 present the first-stage results: a linear probability model of conviction status on judge severity. Standard errors (in parentheses) are clustered at the judge and defendant level, and all continuous variables are standardized to have zero mean and unit variance. In all three regressions, judge severity has a significant effect and the stability of this effect provides evidence of the quasi-random assignment. Nevertheless, we formally test this randomization in column (4), regressing judge severity on case and defendant characteristics and ZIP code fixed-effects. The p-value for the F-statistic of the joint significance of these regressors (including ZIP code effects) is 0.66, providing further support for the exclusion restriction.

In the appendix, Figure 2 shows a local linear logit regression of conviction status on judge severity. It suggests that the monotonicity assumption, which states that defendants who would be convicted by a lenient judge would also be convicted by a more stringent judge, is satisfied.

Column (5) of Table 4 contains the IV regression results for the effect of future criminality V on conviction Y (conditional on covariates). While the statistically insignificant effect of conviction on future criminality is driven partly by the large standard error, the magnitude of the effect is small: conviction decreases future criminality by 0.08 of a standard deviation. It is important to note that a full test of the exclusion restriction requires also conditioning on true guilt Y^*, which is not possible since it is unobserved. However, our IV regression can be interpreted as implicitly accounting for this unobservable Y^* in the following way. The outcome variable V may be influenced by unobserved components, including true guilt Y^*, which are captured in the error term of the IV regression. While Y is likely correlated

Table 4: IV Analysis

	Dependent variable:				
	Convicted			*Judge Severity*	*Future Criminality*
	(1)	(2)	(3)	(4)	(5)
Convicted					−0.08
					(0.19)
Judge Severity	0.05***	0.05***	0.05***		
	(0.001)	(0.003)	(0.003)		
Previous Criminality		0.02***	0.02***	−0.01	0.16***
		(0.003)	(0.003)	(0.04)	(0.02)
White		−0.002	0.01	0.01	−0.16***
		(0.006)	(0.006)	(0.03)	(0.02)
Male		0.008	0.009	−0.003	0.14***
		(0.006)	(0.006)	(0.01)	(0.01)
Violent Crime		−0.19***	−0.19***	−0.11	−0.06
		(0.01)	(0.01)	(0.07)	(0.04)
ZIP Code Fixed-Effects			✓	✓	✓
Observations	29 544	29 544	29 544	29 544	29 544
F-test Statistic	1 616	185	199	0.80	87.1
F-test p-value	0.000	0.000	0.000	0.53	0.000

with Y^* and thus with the error term, we use judge severity as an instrument for Y, the validity of which is given by the random assignment of judges – that is, the severity variable is uncorrelated with the error term.

One further piece of evidence for the conditional mean independence assumption comes from Figure 1 in Section 4. This shows that V does not affect $E[Y|X = x, V]$ for a large range of V in its upper tail. We would expect such plots only if V has no effect on α_1 and α_2, or if V affects α_1 and α_2 only in a certain range of its support — a scenario inconsistent with most conventional functional forms (e.g., linear or quadratic).

7 Conclusion

In this paper, we develop an empirically viable special regressor approach to estimating type I and type II error rates when mistakes are unobserved. Our framework recasts the problem as a misclassified binary choice model and provides new nonparametric identification results that yield simple estimators. Identification relies on two key conditions: an exclusion restriction and a large support assumption. When the exclusion restriction is questionable, we show how the estimands can be interpreted as bounds under a weaker monotonicity condition; we also provide bounds when the large support assumption is violated.

We apply this framework to estimate the probability of a miscarriage of justice in Virginia's court system, carefully examining the assumptions underlying identification, including an IV analysis of whether convictions affect future criminal behavior. While we cannot verify all conditions — particularly the lower-tail large support assumption underlying our wrongful conviction estimates — the results remain informative approximations. Estimates of wrongful convictions can be obtained under additional symmetry assumptions, but these should be treated with caution. In contrast, our diagnostic check suggests that the upper-tail condition is satisfied, allowing estimation of the wrongful acquittal rates: 17–18% for traffic offenses and 32–42% for violent crimes. These figures are conditional on arrest and charge, and should not be interpreted as population-wide error rates. Still, they highlight systematic variation in evidentiary thresholds across offense types and the

procedural realities of high-volume courts.

This framework has applications beyond the judicial context — for example, in medicine, hiring, or algorithmic screening — where unobserved truth and biased decision-making are common. Future work should develop a formal test for the large support condition and explore its applicability in other special regressor models. Inferential procedures for our estimator also remain an open area for research.

8 Disclosure statement

The authors have no conflicts of interest to declare.

9 Data Availability Statement

The data used in this study are publicly available from the Virginia Court Data Project at https://virginiacourtdata.org/. The dataset contains individual-level criminal case records from Virginia's General District and Circuit Courts. All analyses were conducted using only these publicly available data. R code is available at https://github.com/lt9903/R-code-for-papers.

10 Acknowledgments

We thank the Editor, Michal Kolesár, the Associate Editor, and two anonymous referees for their valuable comments, which have greatly improved the paper. We are especially grateful to Professor Kolesár and the Associate Editor for their substantial input, which went well beyond what would normally be expected. Kanaya acknowledges financial support from the Grant-in-Aid for Scientific Research (C, 21K01425) from the Japan Society for the Promotion of Science. This study also received support from the Joint Usage/Research Program of the

Kyoto Institute of Economic Research, Kyoto University. Taylor acknowledges financial support from the Aarhus University Research Fund (AUFF-26852) and the Aarhus Center for Econometrics (ACE), funded by the Danish National Research Foundation (Grant No. DNRF186).

References

Agan, A., Doleac, J. L. & Harvey, A. (2023), 'Misdemeanor Prosecution', *The Quarterly Journal of Economics* **138**(3), 1453–1505.

Andrews, D. W. K. (2000), 'Inconsistency of the Bootstrap When a Parameter Is on the Boundary of the Parameter Space', *Econometrica* **68**(2), 399–405.

Andrews, D. W. K. & Schafgans, M. M. (1998), 'Semiparametric Estimation of the Intercept of a Sample Selection Model', *The Review of Economic Studies* **65**(3), 497–517.

Arnold, D., Dobbie, W. & Hull, P. (2021), 'Measuring Racial Discrimination in Algorithms', *AEA Papers and Proceedings* **111**, 49–54.

Arnold, D., Dobbie, W. & Hull, P. (2022), 'Measuring Racial Discrimination in Bail Decisions', *American Economic Review* **112**(9), 2992–3038.

Bachoc, F., Preinerstorfer, D. & Steinberger, L. (2020), 'Uniformly Valid Confidence Intervals Post-Model-Selection', *Annals of Statistics* **48**(1), 440–463.

Berry, S. T. & Haile, P. A. (2014), 'Identification in Differentiated Products Markets Using Market Level Data', *Econometrica* **82**(5), 1749–1797.

Bjerk, D. & Helland, E. (2020), 'What Can DNA Exonerations Tell Us About Racial Differences in Wrongful-Conviction Rates?', *Journal of Law and Economics* **63**(2), 341–366.

Canay, I. A., Mogstad, M. & Mountjoy, J. (2022), 'On the Use of Outcome Tests for Detecting Bias in Decision Making', *The Review of Economic Studies* **91**(4), 2135–2167.

Claeskens, G. & Van Keilegom, I. (2003), 'Bootstrap Confidence Bands for Regression Curves and Their Derivatives', *Annals of Statistics* **31**(6), 1852–1884.

Fan, J., Heckman, N. E. & Wand, M. P. (1995), 'Local Polynomial Kernel Regression for Generalized Linear Models and Quasi-Likelihood Functions', *Journal of the American Statistical Association* **90**(429), 141–150.

Frölich, M. (2006), 'Non-Parametric Regression for Binary Dependent Variables', *Econometrics Journal* **9**(3), 511–540.

Goh, C. (2018), Rate-Optimal Estimation of the Intercept in a Semiparametric Sample-Selection Model. Working Paper.

Gross, S. R., O'Brien, B., Hu, C. & Kennedy, E. H. (2014), 'Rate of False Conviction of Criminal Defendants Who Are Sentenced to Death', *Proceedings of the National Academy of Sciences* **111**(20), 7230–7235.

Hausman, J. A., Abrevaya, J. & Scott-Morton, F. M. (1998), 'Misclassification of the Dependent Variable in a Discrete-Response Setting', *Journal of Econometrics* **87**(2), 239–269.

Heckman, J. J. & Navarro, S. (2007), 'Dynamic Discrete Choice and Dynamic Treatment Effects', *Journal of Econometrics* **136**, 341–396.

Hull, P. (2021), What Marginal Outcome Tests Can Tell Us About Racially Biased Decision-Making. Working Paper.

Humphries, J. E., Ouss, A., Stavreva, K., Stevenson, M. T. & van Dijk, W. (2023), Conviction, Incarceration, and Recidivism: Understanding the Revolving Door. Forthcoming.

Kamat, V., Norris, S. & Pecenco, M. (2023), Conviction, Incarceration, and Policy Effects in the Criminal Justice System. Working Paper.

Khan, S. & Nekipelov, D. (2018), 'Information Structure and Statistical Information in Discrete Response Models', *Quantitative Economics* **9**(2), 995–1017.

Khan, S. & Tamer, E. (2010), 'Irregular Identification, Support Conditions, and Inverse Weight Estimation', *Econometrica* **78**(6), 2021–2042.

Klein, R. W. & Spady, R. H. (1993), 'An Efficient Semiparametric Estimator for Binary Response Models', *Econometrica* **61**(2), 387–421.

Kolesár, M. (2013), Estimation in an Instrumental Variables Model with Treatment Effect Heterogeneity. Working Paper.

Lewbel, A. (1998), 'Semiparametric Latent Variable Model Estimation with Endogenous or Mismeasured Regressors', *Econometrica* **66**(1), 105–121.

Lewbel, A. (2000*a*), 'Identification of the Binary Choice Model with Misclassification', *Econometric Theory* **16**(4), 603–609.

Lewbel, A. (2000*b*), 'Semiparametric Qualitative Response Model Estimation with Unknown Heteroscedasticity or Instrumental Variables', *Journal of Econometrics* **97**(1), 145–177.

Lewbel, A. & Tang, X. (2015), 'Identification and Estimation of Games with Incomplete Information Using Excluded Regressors', *Journal of Econometrics* **189**(1), 229–244.

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

Magnac, T. & Maurin, E. (2007), 'Identification and Information on Monotone Binary Models', *Journal of Econometrics* **139**(1), 76–104.

Manski, C. F. (1988), 'Identification of Binary Response Models', *Journal of the American Statistical Association* **83**(403), 729–738.

Spencer, B. D. (2007), 'Estimating the Accuracy of Jury Verdicts', *Journal of Empirical Legal Studies* **4**(2), 305–329.

Zalman, M. (2017), 'Wrongful Convictions: A Comparative Perspective', *Journal of Contemporary Criminal Justice* **33**(1), 1–7.