Quantifying Deterrence Effects in Judge Designs

Ian Pitman

September 8, 2024

Abstract

To what extent can judges deter pretrial misconduct without detention? This question is becoming increasingly salient as jurisdictions across the United States enact bail reforms that make most people ineligible for monetary bail and pretrial detention. In this paper, I develop new econometric methods to learn about the deterrence effects of New York City's supervised release program, a *de facto* substitute for monetary bail following bail reform in New York State. I focus on identifying and estimating a simple causal parameter: on average, how much does supervised release increase the court appearance rate of individuals with a given set of characteristics? I propose a parsimonious model of judge behavior that places intuitive restrictions on this parameter while allowing judges to have arbitrary private information. I demonstrate that under certain conditions, these restrictions still hold even if judges make decisions that are noisy, or based on inaccurate beliefs. Finally, I present preliminary empirical results indicating that, on average, pretrial supervision increases court appearance rates by at least 2.9% but no more than 9.9% among the individuals in my sample.

1 Introduction

When judges set bail conditions, they face a dual mandate: maintaining defendants' liberty while ensuring their subsequent appearance in court. Nevertheless, at any given time there are around half a million pretrial detainees in jails across the United States (BJS, 2023). Most of these people are in jail because they have not paid the amount of monetary bail required to secure their release (CCR, 2022). To reduce the rate of pretrial detention, several states including California, New Mexico, Nebraska, Illinois, Indiana, Kentucky, New Jersey, and New York have enacted bail reforms that severely limit, or even eliminate, judges' discretion to set monetary bail. As a result, judges' ability to satisfy their dual mandate hinges on the extent to which they can deter defendants from failing to appear in court without detaining them.

I study the deterrence effects of New York City's supervised release (SR) program, a non-monetary condition that has seen widespread use following New York State's 2020 bail reform. As discussed in greater detail in Section 2.1, program enrollees must regularly check in with a social worker and receive court date reminders via phone or text. There are several features of this empirical context that make it particularly well-suited for studying deterrence effects. First, according to New York State law, judges must only consider failure to appear risk (rather than, say, public safety risk) when making bail decisions. So, the SR program was specifically designed to increase court appearance rates, rather than to decrease recidivism rates more generally. Second, New York State's 2020 bail reform made the majority of defendants ineligible for monetary bail and pretrial detention; judges must effectively decide whether to release these defendants under no conditions or release them under supervision. Eliminating the outside option of detention prevents deterrence effects from being confounded by selection into release, a significant challenge faced by other papers in this literature (Albright, 2022; Rivera, 2023). Third, New York City's SR program is one of the largest and most well-established programs of its kind in the country, acting as a model for similar programs in other states (Mayor's Office of Criminal Justice, 2024). Developing a better understanding of its strengths and weaknesses could help inform bail reform efforts nationwide.

Identifying deterrence effects from observational data requires overcoming the problem of selection: judges will only assign SR to those who they believe are most suitable for the program. The "judge design" represents a canonical solution to this problem in which the researcher runs Two-Stage Least Squares (TSLS) using quasi-random judge assignment as an instrument for supervision (Leslie and Pope, 2017; Dobbie et al., 2018).[1] Unfortunately, it is difficult to justify the usual assumptions from Imbens and Angrist (1994) under which the resulting estimand represents a local average treatment effect. In particular, monotonicity requires all judges to act as if they agree on how defendants should be ranked in terms of their suitability for supervision; in the language of Vytlacil (2002), they must share a common latent index. By ruling out variation in skill across judges, the standard judge design attributes any differences in judge behavior to variation in preferences. In an influential paper, Chan et al. (2022) show that this approach can produce highly misleading results when differences in judge behavior do in fact arise from variation in skill. In Section 3, I present evidence that the concerns raised by Chan et al. (2022) appear to be relevant in my empirical context: canonical TSLS estimates suggest that supervision substantially reduces compliers' court appearance rate, a conclusion that is hard to square with institutional details of the NYC pretrial system. Given that there is no sound theoretical or empirical justification for assuming

[1] Most papers in this literature consider pretrial detention, rather than supervision, as the treatment of interest.

that all NYC judges have the same level of skill, plausible identification of deterrence effects requires new econometric techniques that do not rely on monotonicity.

In the current draft, I focus on identifying conditional average treatment effects (CATEs) of SR on court appearance relative to release under no conditions. Conditioning on observable defendant and court characteristics allows me to concentrate on the fundamental identification challenge: judges may assign supervision based on private information that is both unobserved by the econometrician and heterogeneous across judges. In Section 4.1, I propose a simple behavioral model in which judges trade off the benefit of increased court appearance due to supervision against their perceived cost of supervision, which may vary flexibly across observables but must not depend on unobservables. The importance of ruling out arbitrary unobserved costs in models of treatment selection has been emphasized in the context of detecting bias in judge decision making (Canay et al., 2023). I show that it is also necessary in order to meaningfully restrict judge behavior when monotonicity does not hold. In particular, when judges may have arbitrary and heterogeneous private information, the *only* restriction implied by my behavioral model is positive selection on gains: for any given judge, the defendants she chooses to supervise must have larger treatment effects, on average, than those she doesn't. This restriction is not only useful for tightening the sharp bounds on CATEs, but also robust to certain generalizations of my behavioral model that allow for inaccurate beliefs and idiosyncratic unobserved costs.

This paper builds on a growing literature that has grappled with the question of what can be learned from quasi-random judge assignment without invoking monotonicity. Frandsen et al. (2023) take a reduced-form approach, arguing that Two-Stage Least Squares still recovers a positively weighted average of treatment effects under a weaker "average monotonicity" condition. Unlike the usual monotonicity condition, I am not aware of any equivalence result relating this condition to primitives of a structural model of judge decision making (Vytlacil, 2002). It is difficult to know what restrictions this condition places on the relationship between judge skill and judge preferences, and whether those restrictions are reasonable. My proposed method places no such restrictions, extracting identifying power from any variation in judge skill or judge preferences. Arnold et al. (2022) suggest a "model-free" approach that involves extrapolating conditional average counterfactuals under SR (ROR) from judges who supervise almost all (no) defendants with a given set of characteristics. Unfortunately, for many groups of defendants there may not exist judges with both very high and very low supervision propensities. So, this method will tend to rely heavily on statistical extrapolation which, by design, has little theoretical foundation. By contrast, my proposed method may produce tight bounds on conditional average counterfactuals without placing any requirements on the support of judges' supervision propensities. Both Chan et al. (2022) and Arnold et al. (2022) propose structural models that parameterize the distribution of judges' private information. Both papers adopt an empirical Bayes approach, further parameterizing a prior distribution over judge skill and judge preferences. I demonstrate that it is possible to learn about CATEs without making such strong parametric assumptions, which can be difficult to justify.

Finally, the work of Rambachan (2024) deserves special attention considering its similarity to the current work: the author uses New York City pretrial bail data from before the 2020 bail reform and allows judges to have arbitrary private information. However, he seeks to answer a fundamentally different research question: under what conditions is it possible to establish that judges are making systematic prediction mistakes? Like much of the existing economics literature on pretrial bail, Rambachan (2024) models judges as deciding whether or not to detain defendants based on predictions of whether they would fail to appear in court if released. To assess whether or not these predictions are based on inaccurate beliefs,

the author must circumvent the fact that variation in beliefs may be indistinguishable from variation in preferences under a sufficiently flexible behavioral model. He does this by assuming that preferences do not vary across certain dimensions of defendant characteristics. I make no such assumption, since it is not necessary to distinguish between variation in beliefs and preferences in order to learn about CATEs. Indeed, this flexibility is precisely what makes my results robust to allowing for certain types of inaccurate beliefs. This is particularly important given that Rambachan (2024) documents the existence of inaccurate beliefs in a similar empirical context under the assumptions of his behavioral model.

2 Empirical Setting

2.1 Background on the NYC Pretrial System

Shortly after an individual is arrested in NYC, they appear before a judge at an arraignment hearing. The judge decides what bail conditions the individual must abide by as they await their future court dates. According to New York State law, these conditions must reflect "the kind and degree of control or restriction necessary to reasonably assure the [individual]'s return to court" (New York State Senate, 2023a). To this end, the judge has four main categories of bail conditions at her disposal. First, she may release the defendant under no conditions other than a promise to return to court: this is referred to as release on recognizance (ROR). Second, she may release the defendant under non-monetary conditions, such as mandatory enrollment in a supervised release program. Third, she may set an amount of monetary bail that the defendant must pay in order to be released.[2] Finally, she may detain the defendant outright, ensuring their court appearance at the expense of their pretrial liberty.

Historically, NYC judges have had broad discretion to determine which type of bail condition is appropriate for a given defendant. In practice, this meant that the vast majority of defendants were either ROR or had monetary bail set. For example, in 2018 72% of defendants were ROR while 25% had monetary bail set (DCJS, 2018). Among the latter group, 69% remained in pretrial detention at least 5 days after arraignment as a consequence of not paying bail. This practical reality has lead several authors in the economics literature on pretrial bail to consider monetary bail as a form of de facto detention (Kleinberg et al., 2017; Arnold et al., 2022). According to this view, bail decisions are effectively decisions about whether or not to detain a defendant.

As part of a wave of bail reform efforts across the US, New York State legislators passed the Bail Elimination Act of 2019, restricting judges' discretion to set bail conditions that may result in pretrial detention. Effective January 1st 2020, only defendants charged with a qualifying offense were eligible for monetary bail or outright detention; all other defendants must be ROR or released under non-monetary conditions (New York State Senate, 2023b). Non-qualifying offenses include most misdemeanor and non-violent felony charges, comprising roughly 73% of cases heard in 2022 (OCA, 2024). Defendants charged with non-qualifying offenses represent the population of interest in this paper. Since these defendants are ineligible for detention, they are ideally suited for studying the extent to which judges can increase court appearance rates without detention. While in principle judges may still choose from a variety of non-monetary conditions, in practice SR is by far the most commonly chosen bail condition other than ROR.

[2]The court keeps this money as collateral: the defendant gets it back if they return to court, but not if they don't. There are several types of monetary bail available in New York City, including cash bail and insurance company bail bonds.

Supervised release began in 2009 as a pilot program in Queens to connect defendants with social services, keep them in regular contact with social workers, and remind them of their court dates. According to the government office responsible for administering the program, "the purpose of SR is to help ensure a person's return to court, which is the primary criteria used by judges when making bail decisions under New York State law" (Mayor's Office of Criminal Justice, 2024). Once enrolled in SR, defendants receive court date reminders via phone or text and are required to check in with a social worker between one and four times per month, depending on the level of supervision. Social workers may connect defendants to mental health and substance abuse treatments, or help them apply for public benefits and jobs. Importantly, features of the SR program that may be tailored to a specific defendant, such as the level of supervision, are determined by pretrial service agents, rather than judges. At the arraignment hearing, judges simply decide whether or not to forward defendants to a pretrial service agent responsible for SR enrollment. As of 2018, only 2% of defendants in NYC were assigned SR due to strict eligibility requirements (DCJS, 2018). However, following New York's 2020 bail reform all defendants became eligible for SR, which quickly became a *de facto* substitute for monetary bail in NYC. By 2022, 20% of defendants were assigned SR, while only 10% had monetary bail set (OCA, 2024). The SR program's rapid expansion has come at a significant cost to taxpayers: between 2019 and 2022, the city spent $201 million on its SR contracts (Katz, 2023). This raises important policy questions: has the SR program achieved its objective of increasing court appearance rates? If so, can it be better targeted towards those individuals for whom it is most effective?

2.2 Data and Sample Restrictions

I observe all arraignment hearings in NYC held between January 1st 2020 and January 1st 2024 (OCA, 2024). Each observation includes the following variables:

1. The time and place at which the arraignment occurred, denoted by T. This includes the court, year, and month of the year. Ideally T would also include the courtroom, day of the week, and shift.[3]
2. Defendant characteristics, denoted by C. This includes their race, age, and sex, along with their prior criminal history and the current charge, which can be used to determine if they are charged with a qualifying offense.
3. The identity of the arraignment judge, denoted by the discrete random variable Z on $\{0, \ldots, k\}$.
4. An indicator for whether the defendant was assigned SR, denoted by D.
5. An indicator for whether the defendant subsequently appeared in court, denoted by Y. I infer court appearance if there was no bench warrant issued for failure to appear in court.[4]

To construct my sample of interest, I first keep cases held over three New York City fiscal years from July 1st 2020 to June 30th 2023. This drops cases held during the early months of the COVID-19 pandemic and recent cases that may still be pending.[5] Second, I drop cases that were disposed at arraignment. A

[3] I have not yet received these additional variables from the Office of Court Administration.

[4] I define my outcome of interest to be court appearance Y, rather than failure to appear $1 - Y$. This reframes deterrence effects as incentive effects, providing a more natural link to Roy models of treatment selection (Roy, 1951).

[5] Supervised release was temporarily suspended due to the COVID-19 pandemic between March and June of 2020.

judge will only decide to dispose a case at arraignment if it contains clear legal defects, so in principle this decision should not vary across judges (Leslie and Pope, 2017). Third, I drop defendants who were already incarcerated due to a different offense. Fourth, I drop defendants who were issued a Desk Appearance Ticket (DAT) rather than being taken into custody prior to arraignment. This is because within a given court in a given month, judge assignment may differ systematically based on whether a defendant was issued a DAT or not. Fifth, I drop Kings Criminal Court due to strong evidence of non-random judge assignment in a given month even after dropping defendants with DATs.[6] Sixth, I drop defendants charged with a qualifying offense according to New York State Criminal Procedure Law (New York State Senate, 2023b). Finally, for the sake of precision I drop court-by-judge bins containing fewer than 30 cases over the three year sampling period.

2.3 Counterfactuals and Preliminary Assumptions

Let $Y(0)$ and $Y(1)$ denote counterfactual court appearance under ROR and SR, respectively, and let $D(z)$ denote counterfactual SR assignment under judge z. Observables are related to counterfactuals by the following switching equations

$$D = D(Z) = \sum_{z=0}^{k} \mathbb{1}\{Z = z\}D(z), \qquad Y = Y(D) = (1 - D)Y(0) + DY(1)$$

Implicit in this notation is the usual exclusion restriction: judges must only affect court appearance rates through SR assignment. Importantly, while judges cannot set monetary bail or detain defendants with non-qualifying offenses, they may set other non-monetary conditions instead of or in addition to SR. While in principle this represents a threat to exclusion, in practice judges rarely exercise this option. Among defendants with non-qualifying offenses, 14% are assigned SR while the next most common non-monetary conditions are "No Firearms or Weapons" (0.8%) and "Obey Order of Protection" (0.02%). Unlike SR, these conditions are usually intended to protect the safety of certain individuals (e.g. victims of domestic violence) rather than increase court appearance rates. Given these institutional details, the exclusion restriction appears plausible in this setting.

Following the literature on pretrial bail in New York City, I assume that judges are quasi-randomly assigned to defendants conditional on the court, year, month, and day of the week (Kleinberg et al., 2017). This comes from the fact that judges are assigned to arraignment shifts by a rotating calendar system.

Assumption 1 (Conditional Exogeneity)**.**

$$(Y(0), Y(1), D(0), \ldots, D(k), C) \perp\!\!\!\perp Z \mid T$$

Several papers have verified that defendant characteristics appear to be unrelated to judge assignment conditional on T using data from New York City before the 2020 bail reform (Kleinberg et al., 2017; Arnold et al., 2022). Defining $X \coloneqq C, T$, Assumption 3 implies that $(Y(0), Y(1), D(0), \ldots, D(k)) \perp\!\!\!\perp Z \mid X$, the relevant exogeneity condition for the identification results in Section 4.

A key feature of the pretrial bail setting is that bail conditions are naturally ordered in terms of restrictiveness. This can be seen, for example, in the legal requirement that judges consider "the... degree of... restriction necessary" to ensure court appearance when making bail decisions (New York State

[6] In principle, this could be remedied by conditioning on additional variables such as the courtroom, day of the week, and shift. Unfortunately, I have not yet received these variables from the Office of Court Administration.

Senate, 2023a). Implicit in this language is the notion that, *ceteris paribus*, a defendant will be more likely to appear in court under a more restrictive condition relative to a less restrictive one. Motivated by this idea, I assume that SR weakly increases court appearance relative to ROR.

Assumption 2 (Monotone Treatment Response)**.**

$$\Pr(Y(1) \geq Y(0)) = 1$$

Put another way, I assume that there are no defendants who would appear in court if assigned ROR, but would not if assigned SR. Interventions implemented by the SR program are designed to overcome barriers preventing defendants from appearing in court, including inattentiveness, addiction, and financial insecurity. Under Assumption 2, while these interventions might not work (we can have $Y(0) = Y(1) = 0$) they cannot produce the opposite of their intended effect. One possibility that would cast doubt on Assumption 2 is if bench warrants were issued for noncompliance with SR conditions (such as failure to check in with a social worker), rather than failure to appear in court. Rivera (2023) documents this phenomenon, in which violations of non-monetary conditions themselves are charged as crimes, in the case of electronic monitoring in Chicago. Unfortunately, my data do not allow me to rule out this possibility directly. However, according to court administrators and SR staff members, there are rarely any negative consequences for noncompliance with SR so long as the defendant appears for their required court dates (MDRC, 2020).

3 Empirical Results using Two-Stage Least Squares

Before specifying a model of judge behavior, I will provide evidence that canonical judge design methodologies are not well-suited to the empirical setting at hand. In particular, the usual Imbens and Angrist (1994) monotonicity condition as well as the weaker notion of average monotonicity proposed by Frandsen et al. (2023) ensure that the TSLS estimand recovers a positively weighted average of treatment effects.[7] In this case, we would expect the TSLS estimand to be non-negative, since treatment effects are non-negative under Assumption 2. However, TSLS and jackknife instrumental variables estimates of the

[7]Putting aside the issues of misspecification raised by Blandhol et al. (2022).

effect of supervision on court appearance reported in Table 1 are negative and significant at the 5% and 10% levels, respectively.

Table 1: OLS, TSLS, and IJIVE Estimated Effects of SR on Court Appearance and Recidivism

	OLS	TSLS	IJIVE	OLS	TSLS	IJIVE
Court Appearance	-0.116^{***}	-0.069^{**}	-0.063^{*}	-0.069^{***}	-0.058^{*}	-0.059^{*}
	(0.005)	(0.030)	(0.036)	(0.005)	(0.030)	(0.035)
Recidivism	0.184^{***}	0.056	0.032	0.093^{***}	0.028	0.017
	(0.006)	(0.040)	(0.048)	(0.006)	(0.039)	(0.046)
Defendant Characteristics	No	No	No	Yes	Yes	Yes
Court Appearance Rate	0.922	0.922	0.922	0.922	0.922	0.922
Recidivism Rate	0.155	0.155	0.155	0.155	0.155	0.155
Number of Judges	167	167	167	167	167	167
Number of Cases	75,262	75,262	75,262	75,262	75,262	75,262

Notes: This table reports estimated coefficients on an indicator for supervised release, the treatment of interest, across several regression specifications. Rows correspond to different outcome variables, while columns correspond to different regression procedures. All regressions control for court by year month fixed effects T, while regressions in the last three columns also control for defendant characteristics C. Standard errors are heteroskedasticity robust but not clustered, since the current dataset does not contain the required clustering variables. Stars denote $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

The regression specifications in Table 1 adhere to current best practices in judge designs as outlined by Chyn et al. (2024). Following existing literature, I also include results using an indicator for recidivism as the outcome of interest. OLS estimates suggest that selection bias is severe in this empirical context: defendants assigned SR are much less likely to appear in court and much more likely to recidivate. Including defendant characteristics substantially reduces the magnitude of these estimates, indicating that certain characteristics are strong confounders. I produce TSLS estimates by using a full set of judge assignment indicators to instrument for SR. This design may suffer from many-instruments bias, which I eliminate using the improved jackknife procedure (IJIVE) proposed by Ackerberg and Devereux (2009).[8] Both TSLS and IJIVE yield similar conclusions: namely, that pretrial supervision reduces compliers' court appearance rate by roughly 6%.

There are several possible explanations for why TSLS recovers the "wrong sign", in the sense that it is not consistent with Assumption 2. The first is simply that Assumption 2 does not hold, meaning judges could increase court appearance rates by choosing not to supervise marginal defendants. As discussed in Sections 2.1 and 2.3, this explanation is difficult to reconcile with institutional details of the NYC pretrial system: judges' objective is to increase court appearance rates by assigning supervision, an intervention specifically designed for this purpose.

A more plausible explanation might be that Assumption 1 does not hold. For example, in a given court on a given month, judges with higher supervision propensities might tend to be assigned defendants with lower court appearance rates. It is difficult to rule out this possibility in the current draft, since I do not yet have access to all of the conditioning variables T typically used in the literature on NYC

[8]The IJIVE procedure corresponds closely to specifications widely adopted in the literature in which a residualized leave-out-mean measure of judge SR propensity is used as the instrument. The unbiased jackknife estimator (UJIVE) proposed by Kolesár (2013) yields nearly identical results.

pretrial bail decisions to justify Assumption 1. However, Appendix Table 3 presents a test of exogeneity in which first-stage fitted values from a TSLS / IJIVE specification *without* defendant characteristics C are regressed on C and T. If $Z \perp\!\!\!\perp C \mid T$, the coefficients on C in this regression should be zero. An F-test of the joint nullity of these coefficients fails to reject for the IJIVE specification, with a p-value of 0.17. This lends some credence to the notion that judges are quasi-randomly assigned conditional on T, despite the limitations of the current dataset.

A final explanation would involve failures of either exclusion or monotonicity. Indeed, the joint test of exclusion and monotonicity proposed by Frandsen et al. (2023) rejects at the 1% level for all of the specifications considered in Table 1. Even if we replace these assumptions with their "average" counterparts from Frandsen et al. (2023), Assumption 2 still guarantees a non-negative TSLS estimand. By contrast, Chan et al. (2022) propose a plausible mechanism through which monotonicity violations may cause TSLS to recover the wrong sign, a phenomenon that they document in the context of radiologists making pneumonia diagnoses. Even if supervision can only weakly increase court appearance, a less skilled judge may choose to supervise more defendants than a more skilled judge but nevertheless achieve a lower court appearance rate. This would produce an inverse relationship between supervision rates and court appearance rates across judges, resulting in a negative TSLS estimand.[9] Ultimately, TSLS fails in this example because it does not account for variation in the quality of judges' signals about potential outcomes. This motivates developing a model of judge behavior that incorporates heterogeneous signal quality, allowing the researcher to learn about treatment effects not only from variation in judge preferences but also from variation in judge skill.

4 Behavioral Model

I model judges as facing a decision problem with the following structure

1. A defendant appears before judge z at an arraignment hearing. Judge z's information set consists of X, S_z, where X includes court-by-time and defendant characteristics observed by all judges and S_z is an exogenous judge-specific signal. X is observed by the econometrician, while S_z is not.[10]

2. For $d \in \{0, 1\}$, judge z's *ex post* utility of assigning release condition $D(z) = d$ is given by $\mathcal{U}_z(d)$, which may depend on X, S_z and the defendant's counterfactual court appearance $Y(0), Y(1)$. Judge z uses her information set X, S_z to form a posterior $\mathcal{P}_z$ over the unobserved state $Y(0), Y(1)$ given her beliefs.

3. Judge z chooses $D(z)$ to maximize *ex ante* expected utility $\mathcal{E}_z[\mathcal{U}_z(D(z)) \mid X, S_z]$, where $\mathcal{E}_z$ denotes her expectation with respect to the posterior $\mathcal{P}_z$.

4. Finally, the defendant's court appearance $Y(D(z))$ is realized, along with judge z's *ex post* utility $\mathcal{U}_z(D(z))$.

There are three model primitives that characterize this decision problem: judges' expectation operators $\mathcal{E}_z$, preferences $\mathcal{U}_z$, and signals S_z. In Section 4.1, I propose a baseline model in which

[9] For a visual representation of this phenomenon, see Panel B of Figure 1

[10] I sometimes refer to judges' signals S_z as their private information. This information is only "private" in the sense that it is not observed by the econometrician. Although all judges are assumed to observe X, I do not take a stand on what components of S_z may or may not be shared across judges.

I assume rational expectations and adopt preferences from the Extended Roy Model (Heckman and Vytlacil, 2007; D'Haultfœuille and Maurel, 2013). However, unlike the Extended Roy Model, I allow for unrestricted heterogeneity in signals S_z across judges, and make no assumptions regarding their distribution. In Section 4.2, I show that under the baseline model, conditional average counterfactuals $\mathbb{E}[Y(0) \mid X = x], \mathbb{E}[Y(1) \mid X = x]$ must satisfy a simple restriction which, along with Assumptions 1 and 2, characterizes their sharp identified set. Finally, in Sections 4.3 and 4.4 I relax the baseline model's strong assumptions on expectations and preferences and provide sufficient conditions under which my main identification results from Section 4.2 still hold.

4.1 Baseline Model

Assumption 3 (Extended Roy Model with Heterogeneous Signals)**.** *For all* $z \in \{0, \ldots, k\}$,

$$Y(0), Y(1) \perp\!\!\!\perp D(z) \mid X, S_z \tag{1}$$

$$D(z) \in \arg\max_{d \in \{0,1\}} \mathbb{E}[\mathcal{U}_z(d) \mid X, S_z] \tag{2}$$

$$\mathcal{U}_z(d) = Y(d) - \lambda_z(X)d \tag{3}$$

Equation (1) formalizes the sense in which X, S_z constitutes judge z's information set. It is only meaningful if $D(z)$ is a non-degenerate random variable conditional on X, S_z. In this case, Equation (2) implies that judge z must be indifferent between ROR and SR *ex ante*. So, any tie-breaking mechanism must be independent of the unobserved state. If not, then judge z must have received information about the unobserved state that is not captured by X, S_z, contradicting the notion that X, S_z represents her information set.

Equation (2) posits rational expectations, replacing judges' expectation operators $\mathcal{E}_z$ with the expectation operator of the underlying data generating process $\mathbb{E}$. In other words, judges are assumed to have accurate beliefs about the distribution of $Y(0), Y(1)$ and update those beliefs according to Bayes' rule. In Section 4.3, I relax this assumption by allowing $\mathcal{E}_z$ to arise from Bayesian updating of inaccurate beliefs conditional on X.

Equation (3) specifies that the *ex post* net benefit of assigning SR relative to ROR is given by

$$\mathcal{U}_z(1) - \mathcal{U}_z(0) = Y(1) - Y(0) - \lambda_z(X)$$

Judges weigh the benefit of increased court appearance due to supervision, $Y(1) - Y(0)$, against their perceived cost of assigning supervision, $\lambda_z(X)$. This cost is measured in units of court appearance, so it represents judge z's marginal rate of substitution between supervision and court appearance. Differences in costs across judges may arise for a variety of reasons. For example, if judge z assesses the negative welfare impact of supervision for defendants with characteristics $C \subset X$ to be large relative to judge z', we may have $\lambda_z(X) > \lambda_{z'}(X)$, *ceteris paribus*. Alternatively, if judge z considers the cost of a missed court date to be large relative to judge z', we may have $\lambda_z(X) < \lambda_{z'}(X)$, *ceteris paribus*. Importantly, although costs may be heterogeneous across judges, costs for a given judge must only depend on observables. This is what distinguishes preferences in an Extended Roy Model from those in a Generalized Roy Model, which also allows for unobserved costs (Heckman and Vytlacil, 2007). In Section 4.4, I relax this requirement by incorporating idiosyncratic taste shocks, a particular type of unobserved cost.

To illustrate the empirical content of Assumption 3, I will begin by defining judge z's marginal treatment effect (MTE) curve conditional on $X = x$. Denote judge z's *ex ante* expected gross benefit

of supervision $\mathbb{E}[Y(1) - Y(0) \mid X = x, S_z]$ by V_z, and suppose for the sake of exposition that its cdf $F_{V_z|X=x}$ is continuous.[11] Then by Assumption 3, among defendants with $X = x$

$$\begin{aligned} D(z) &= \mathbb{1}\{\mathbb{E}[Y(1) - Y(0) - \lambda_z(X) \mid X = x, S_z] \geq 0\} \\ &= \mathbb{1}\{V_z \geq \lambda_z(x)\} \\ &= \mathbb{1}\{F_{V_z|X=x}(V_z) \geq F_{V_z|X=x}(\lambda_z(x))\} \\ &= \mathbb{1}\{\underbrace{1 - F_{V_z|X=x}(V_z)}_{:= U_z} \leq 1 - F_{V_z|X=x}(\lambda_z(x))\} \end{aligned}$$

where we have taken the probability integral transform of V_z to arrive at a uniformly distributed latent index U_z. Judge z's MTE curve is given by the following function of $u \in [0, 1]$

$$\begin{aligned} \text{MTE}_z(u \mid x) &:= \mathbb{E}[Y(1) - Y(0) \mid X = x, U_z = u] \\ &= \mathbb{E}[Y(1) - Y(0) \mid X = x, 1 - F_{V_z|X=x}(V_z) = u] \\ &= \mathbb{E}[Y(1) - Y(0) \mid X = x, V_z = F^{-1}_{V_z|X=x}(1 - u)] \\ &= \mathbb{E}[\mathbb{E}[Y(1) - Y(0) \mid X = x, V_z = F^{-1}_{V_z|X=x}(1 - u), S_z] \mid X = x, V_z = F^{-1}_{V_z|X=x}(1 - u)] \\ &= \mathbb{E}[\mathbb{E}[Y(1) - Y(0) \mid X = x, S_z] \mid X = x, V_z = F^{-1}_{V_z|X=x}(1 - u)] \\ &= \mathbb{E}[V_z \mid X = x, V_z = F^{-1}_{V_z|X=x}(1 - u)] \\ &= F^{-1}_{V_z|X=x}(1 - u) \end{aligned}$$

Notice that $\text{MTE}_z(u \mid x)$ is positive (by Assumption 2) and decreasing (since $F^{-1}_{V_z|X=x}$ is increasing). By implementing a threshold rule on U_z, judge z only assigns supervision to those defendants with the highest *ex ante* expected treatment effects of supervision. This means that for any given rate of supervision r, judge z achieves the highest feasible court appearance rate given her signal. We will refer to this function of $r \in [0, 1]$ as her MTE frontier

$$\mathcal{F}_z(r \mid x) := \mathbb{E}[Y(0) \mid X = x] + \int_0^r \text{MTE}_z(u \mid x) \mathrm{d}u$$

The function $\mathcal{F}_z(r \mid x)$ is increasing and concave, and can be seen as a production–possibility frontier in supervision by court appearance rate space. By Assumption 3, judge z's *ex ante* expected utility conditional on $X = x$ is given by

$$\begin{aligned} \mathbb{E}[\mathcal{U}_z(D(z)) \mid X = x] &= \mathbb{E}[Y(D(z)) - \lambda_z(x)D(z) \mid X = x] \\ &= \mathbb{E}[Y \mid X = x, Z = z] - \lambda_z(x)\mathbb{E}[D \mid X = x, Z = z] \quad \text{by Assumption 1} \end{aligned}$$

So, Assumption 3 admits a simple economic interpretation: we can view judge z as directly choosing her supervision and court appearance rates to maximize her expected utility subject to the feasibility constraint imposed by her MTE frontier

$$\begin{aligned} \mathbb{E}[D \mid X = x, Z = z], \mathbb{E}[Y \mid X = x, Z = z] \in \underset{r,y \in [0,1]\times[0,1]}{\arg\max} \quad & y - \lambda_z(x)r \\ \text{s.t.} \quad & y \leq \mathcal{F}_z(r \mid x) \end{aligned}$$

[11] This avoids the possibility that indifference occurs with positive probability, $\Pr(V_z = \lambda_z(x) \mid X = x) > 0$.

The constraint binds at the optimum and, if $\mathbb{E}[D \mid X = x, Z = z] \in (0, 1)$,

$$\lambda_z(x) = \left.\frac{\partial \mathcal{F}_z(r \mid x)}{\partial r}\right|_{\mathbb{E}[D|X=x,Z=z]} = \text{MTE}_z(\mathbb{E}[D \mid X = x, Z = z] \mid x)$$

so the marginal cost of supervision is equal to the benefit of assigning supervision to a marginal defendant.

Assumption 3 is considerably weaker than the canonical Extended Roy Model, in which it is typically assumed that all judges receive the same signal $S_z := S$ (Canay et al., 2023). As shown by Vytlacil (2002), this implies the monotonicity assumption of Imbens and Angrist (1994), and is a sufficient condition for all judges to have the same MTE frontier.[12] By shutting down heterogeneity in beliefs and signals, the Extended Roy Model attributes any variation in behavior across judges to heterogeneity in preferences.[13] As shown by Chan et al. (2022), this can produce highly misleading results when variation in judge behavior is in fact driven by heterogeneity in signals rather than preferences.

For example, suppose two judges z and z' have the same preferences $\lambda_z(x) = \lambda_{z'}(x)$ but V_z is a mean-preserving spread of $V_{z'}$. Judge z is more skilled than judge z' in the sense that S_z is Blackwell more informative than $S_{z'}$ about the payoff-relevant unobserved state $Y(1) - Y(0)$ (Blackwell, 1953). Figure 1 illustrates two possibilities for how this could translate into differences in judge behavior.

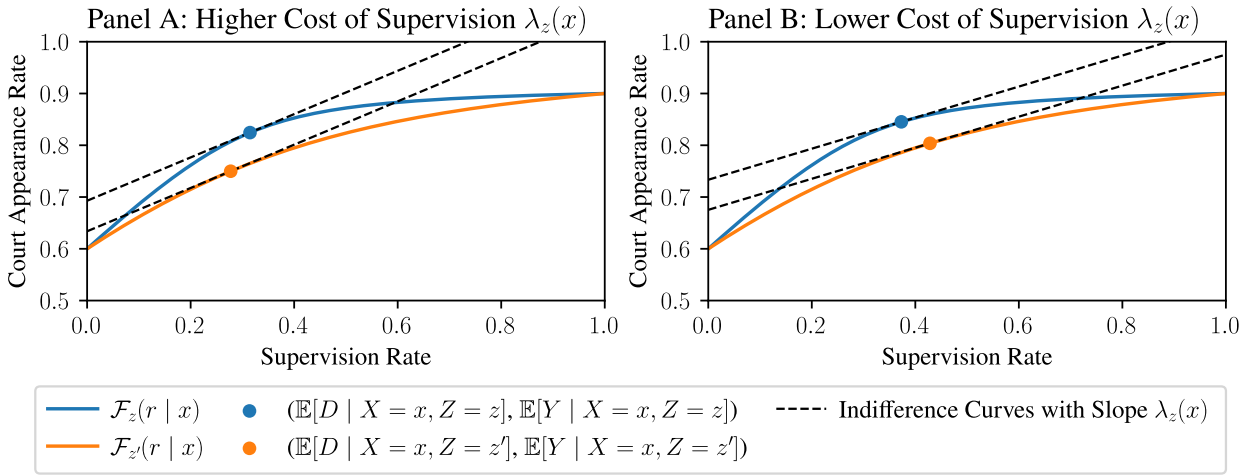

Figure 1: Judge Behavior Across Skill Levels, Holding Preferences Fixed

Notice that in both cases, the Wald estimand

$$\frac{\mathbb{E}[Y \mid X = x, Z = z'] - \mathbb{E}[Y \mid X = x, Z = z]}{\mathbb{E}[D \mid X = x, Z = z'] - \mathbb{E}[D \mid X = x, Z = z]}$$

is outside $[0, 1]$, the support of treatment effects; in Panel A it is equal to 2, while in Panel B it is equal to -0.75. Consequently, it cannot possibly represent a positively weighted average of treatment effects. Furthermore, Figure 1 shows that the sign of the bias is indeterminate and may be sensitive to small changes in judge preferences. So, despite Assumption 3 holding in this simple example, methods such as

[12] In principle, judges may receive different signals so long as they all produce the same distribution $F_{V_z|X=x}$. These signals are equivalent from the perspective of expected utility maximization, since they generate the same distribution of posteriors over the payoff-relevant unobserved state $Y(1) - Y(0)$.

[13] As discussed in Section 4.3, certain types of heterogeneity in beliefs are indistinguishable from heterogeneity in preferences, and can therefore be incorporated into the Extended Roy Model.

Two-Stage Least Squares or Local Instrumental Variables will fail to capture any kind of local average treatment effect (Imbens and Angrist, 1994; Heckman and Vytlacil, 1999).

The notion that violations of monotonicity can threaten the validity of instrumental variables designs has been well-understood ever since montonicity was first proposed (Angrist et al., 1996). However, judge designs provide a particularly plausible setting in which such violations may occur: judges may have different levels of experience, or different mental limitations that affect the quality of their signals. Indeed, recently proposed tests designed to detect situations like Figure 1 have rejected monotonicity across several empirical contexts, including the context of bail decisions in New York City before the 2020 bail reform (Frandsen et al., 2023; Chan et al., 2022; Arnold et al., 2022). As a result, a growing literature has considered the question of what can be learned from judge designs when monotonicity does not hold. In Section 4.2, I contribute to this literature by showing that, despite placing no restrictions on judges' private information, Assumption 3 nevertheless tightens the sharp bounds on conditional average treatment effects.

4.2 Identification Results under the Baseline Model

As discussed in Section 4.1, Assumptions 1 and 3 imply that judges act as if they directly choose supervision and court appearance rates on their MTE frontiers

$$\mathbb{E}[Y \mid X = x, Z = z] = \mathcal{F}_z(\mathbb{E}[D \mid X = x, Z = z] \mid x)$$

in order to maximize their expected utility. For all judges, we know that

$$\mathcal{F}_z(0 \mid x) = \mathbb{E}[Y(0) \mid X = x], \qquad \mathcal{F}_z(1 \mid x) = \mathbb{E}[Y(1) \mid X = x]$$

so they could have chosen to supervise noone (everyone) and achieve the conditional average counterfactual under ROR (SR) as their court appearance rate. By revealed preference, both of these options must not have yielded a higher expected utility than the option they actually chose

$$\mathbb{E}[Y(0) \mid X = x] \leq \mathbb{E}[Y \mid X = x, Z = z] - \lambda_z(x)\mathbb{E}[D \mid X = x, Z = z] \tag{4}$$

$$\mathbb{E}[Y(1) \mid X = x] - \lambda_z(x) \leq \mathbb{E}[Y \mid X = x, Z = z] - \lambda_z(x)\mathbb{E}[D \mid X = x, Z = z] \tag{5}$$

In the case of an interior solution, we can rearrange terms to get

$$\frac{\mathbb{E}[Y(1) \mid X = x] - \mathbb{E}[Y \mid X = x, Z = z]}{1 - \mathbb{E}[D \mid X = x, Z = z]} \leq \lambda_z(x) \leq \frac{\mathbb{E}[Y \mid X = x, Z = z] - \mathbb{E}[Y(0) \mid X = x]}{\mathbb{E}[D \mid X = x, Z = z]}$$

which, by Assumption 1, is equivalent to

$$\begin{aligned}
&\text{ATU}_z(x) \leq \lambda_z(x) \leq \text{ATT}_z(x) \\
&\text{ATU}_z(x) \coloneqq \mathbb{E}[Y(1) - Y(0) \mid X = x, D(z) = 0] \\
&\text{ATT}_z(x) \coloneqq \mathbb{E}[Y(1) - Y(0) \mid X = x, D(z) = 1]
\end{aligned}$$

So, regardless of judges' preferences, it must be the case that $\text{ATU}_z(x) \leq \text{ATT}_z(x)$, meaning there is positive selection on gains from supervision. Alternatively, we can directly deduce the existence of positive selection on gains from the fact that judge-specific MTE curves are decreasing, as derived in Section 4.1.

The central idea behind my identification results is that the implication of positive selection on gains exhausts the empirical content of Assumption 3 regarding conditional average counterfactuals. In

particular, for any pair $\mathbb{E}[Y(0) \mid X = x], \mathbb{E}[Y(1) \mid X = x]$ consistent with Assumptions 1 and 2 that satisfies $\text{ATU}_z(x) \leq \text{ATT}_z(x)$ for all $z \in \{0, \ldots, k\}$, there exist preferences $\{\lambda_z(x)\}_{z=0}^k$ and signals $\{S_z\}_{z=0}^k$ such that judges behave according to Assumption 3. I formalize this idea in the following theorem.

Theorem 1. *Under Assumptions 1, 2 and 3, the sharp identified set for the pair of conditional average counterfactuals* $\mathbb{E}[Y(0) \mid X = x], \mathbb{E}[Y(1) \mid X = x]$ *for any* $x \in \text{supp}(X)$ *is given by the convex polygon*

$$\begin{aligned}
\mathcal{Y}(x) = \Big\{ & y_0, y_1 \in [0,1] \times [0,1] \text{ such that for all } z \in \{0, \ldots, k\}, \\
& y_0 \geq \mathbb{E}[(1-D)Y \mid X = x, Z = z] && (6) \\
& y_0 \leq \mathbb{E}[Y \mid X = x, Z = z] - \mathbb{E}[D \mid X = x, Z = z](y_1 - y_0) && (7) \\
& y_1 \geq \mathbb{E}[Y \mid X = x, Z = z] && (8) \\
& y_1 \leq 1 - \mathbb{E}[D(1-Y) \mid X = x, Z = z] \Big\} && (9)
\end{aligned}$$

which yields the following sharp bounds on the conditional average treatment effect

$$\begin{aligned}
&\underline{\tau}(x) \leq \mathbb{E}[Y(1) - Y(0) \mid X = x] \leq \overline{\tau}(x) \\
&\underline{\tau}(x) \coloneqq \max \Bigg\{ \max_{\tilde{z}} \{\mathbb{E}[Y \mid X = x, Z = \tilde{z}]\} - \min_{z} \{\mathbb{E}[Y \mid X = x, Z = z]\}, \\
&\qquad \max_{z \in \mathcal{Z}_0} \left\{ \frac{\max_{\tilde{z}} \{\mathbb{E}[Y \mid X = x, Z = \tilde{z}]\} - \mathbb{E}[Y \mid X = x, Z = z]}{1 - \mathbb{E}[D \mid X = x, Z = z]} \right\} \Bigg\} \\
&\overline{\tau}(x) \coloneqq \min \Bigg\{ \min_{z} \{1 - \mathbb{E}[D(1-Y) \mid X = x, Z = z]\} - \max_{\tilde{z}} \{\mathbb{E}[(1-D)Y \mid X = x, Z = \tilde{z}]\}, \\
&\qquad \min_{z \in \mathcal{Z}_1} \left\{ \frac{\mathbb{E}[Y \mid X = x, Z = z] - \max_{\tilde{z}} \{\mathbb{E}[(1-D)Y \mid X = x, Z = \tilde{z}]\}}{\mathbb{E}[D \mid X = x, Z = z]} \right\} \Bigg\} \\
&\mathcal{Z}_0 \coloneqq \{z \in \{0, \ldots, k\} : \mathbb{E}[D \mid X = x, Z = z] \neq 1\} \\
&\mathcal{Z}_1 \coloneqq \{z \in \{0, \ldots, k\} : \mathbb{E}[D \mid X = x, Z = z] \neq 0\}
\end{aligned}$$

Inequality 7 is equivalent to $\text{ATU}_z(x) \leq \text{ATT}_z(x)$, the key restriction coming from Assumption 3. Without Assumption 3, the sharp identified set $\mathcal{Y}(x)$ would be defined similarly, but with Inequality 7 replaced by $y_0 \leq \mathbb{E}[Y \mid X = x, Z = z]$. In this case, $\mathcal{Y}(x)$ would be a combination of Manski instrumental variable and Manski monotone treatment response bounds (Manski, 1990, 1997). If, in the spirit of Manski and Pepper (2000)'s monotone treatment selection assumption, Inequality 7 were directly invoked as a model primitive instead of Assumption 3, we would arrive at exactly the same sharp identified set as in Theorem 1. The primary contribution of Theorem 1, then, is to provide a strong decision theoretic foundation for positive selection on gains as the *only* restriction implied by an Extended Roy Model with heterogeneous signals.

Theorem 1 is closely related to fundamental results from the literature on information design and robust predictions: specifically, Theorem 1 of Bergemann and Morris (2016), as operationalized by Gualdani and Sinha (2019) for the case of discrete choice models. Since we have assumed rational expectations, we can consider conditional average counterfactuals as summarizing judges' common prior over the unobserved state $Y(0), Y(1)$ conditional on $X = x$. Assumptions 1 and 2 place restrictions on this common prior in the form of Manski bounds, even in the absence of a behavioral model. If, in addition to these restrictions, Inequality 7 holds, then there exist preferences $\{\lambda_z(x)\}_{z=0}^k$ such that Inequalities 4 and 5 hold. These

inequalities, referred to as "obedience conditions", in turn ensure the existence of signals $\{S_z\}_{z=0}^k$ such that judges behave according to Assumption 3 given their common prior (Bergemann and Morris, 2016).

While preference parameters $\lambda_z(x)$ do not appear in Theorem 1, sharp bounds on $\lambda_z(x)$ in the case of an interior solution follow from the fact that $\text{ATU}_z(x) \leq \lambda_z(x) \leq \text{ATT}_z(x)$.

Corollary 1. *Under Assumptions 1, 2 and 3, sharp bounds on preference parameters $\lambda_z(x)$ are given by*

$$\underline{\lambda}_z(x) \leq \lambda_z(x) \leq \overline{\lambda}_z(x)$$

$$\underline{\lambda}_z(x) \coloneqq \begin{cases} \underline{\tau}(x) & \text{if} \quad \mathbb{E}[D \mid X = x, Z = z] = 0 \\ \frac{\max_{\tilde{z}}\{\mathbb{E}[Y|X=x,Z=\tilde{z}]\} - \mathbb{E}[Y|X=x,Z=z]}{1-\mathbb{E}[D|X=x,Z=z]} & \text{if} \quad \mathbb{E}[D \mid X = x, Z = z] \in (0,1) \\ -\infty & \text{if} \quad \mathbb{E}[D \mid X = x, Z = z] = 1 \end{cases}$$

$$\overline{\lambda}_z(x) \coloneqq \begin{cases} \infty & \text{if} \quad \mathbb{E}[D \mid X = x, Z = z] = 0 \\ \frac{\mathbb{E}[Y|X=x,Z=z] - \max_{\tilde{z}}\{\mathbb{E}[(1-D)Y|X=x,Z=\tilde{z}]\}}{\mathbb{E}[D|X=x,Z=z]} & \text{if} \quad \mathbb{E}[D \mid X = x, Z = z] \in (0,1) \\ \overline{\tau}(x) & \text{if} \quad \mathbb{E}[D \mid X = x, Z = z] = 1 \end{cases}$$

where $\underline{\tau}(x), \overline{\tau}(x)$ are defined as in Theorem 1. In particular, for all $z \in \{0, \ldots, k\}$,

$$\underline{\lambda}_z(x) \leq \underline{\tau}(x) \leq \overline{\tau}(x) \leq \overline{\lambda}_z(x)$$

As discussed in Section 4.1, at an interior solution $\lambda_z(x)$ represents the marginal treatment effect for defendants with $X = x$ and $U_z = \mathbb{E}[D \mid X = x, Z = z]$. In stark contrast to the canonical Extended Roy Model, Assumptions 1, 2 and 3 are always more informative about conditional average treatment effects than they are about marginal treatment effects. In addition, while Theorem 1 is robust to the presence of certain inaccurate beliefs or taste shocks as described in Sections 4.3 and 4.4, Corollary 1 is not.

4.3 Incorporating Inaccurate Beliefs

Assumption 3 specifies that judges have rational expectations. In particular, they have accurate prior beliefs about the unobserved state $Y(0), Y(1)$ conditional on X, denoted by

$$\pi^{y_0 y_1}(x) \coloneqq \Pr(Y(0) = y_0, Y(1) = y_1 \mid X = x)$$

For the sake of exposition, suppose judge signals S_z admit a probability density function conditional on $Y(0)$, $Y(1)$, and X, denoted by

$$f_z^{y_0 y_1}(s \mid x) \coloneqq f_{S_z|Y(0),Y(1),X}(s \mid y_0, y_1, x)$$

Then under Assumption 3, judges update their prior beliefs to form posteriors according to Bayes' rule

$$\Pr(Y(0) = y_0, Y(1) = y_1 \mid X, S_z) = \frac{\pi^{y_0 y_1}(X) f_z^{y_0 y_1}(S_z \mid X)}{\pi^{00}(X) f_z^{00}(S_z \mid X) + \pi^{01}(X) f_z^{01}(S_z \mid X) + \pi^{11}(X) f_z^{11}(S_z \mid X)}$$

where $\pi^{10}(X) = 0$ by Assumption 2.

While the assumption of accurate beliefs is convenient in that it provides a tight link between judges' decision making process and the underlying DGP, it can be difficult to justify in the context of pretrial bail decisions. For example, Rambachan (2024) estimates that more than a fifth of New York City judges before the 2020 bail reform behaved in a manner inconsistent with expected utility maximization at accurate

beliefs. He finds that these judges' beliefs underreact to predictable variation in court appearance rates across observable characteristics, which is consistent with evidence from the literature on algorithmic decision aids (Angelova et al., 2023). This motivates accommodating inaccurate beliefs with the following generalization of Assumption 3.

Assumption 3′ (Assumption 3 with Inaccurate Beliefs)**.** *Replace Equation* (2) *in Assumption 3 by*

$$D(z) \in \arg\max_{d \in \{0,1\}} \mathcal{E}_z[\mathcal{U}_z(d) \mid X, S_z]$$

where $\mathcal{E}_z$ *is a subjective expectation with respect to the posterior* $\mathcal{P}_z$ *over* $Y(0), Y(1)$ *arising from Bayesian updating of potentially inaccurate beliefs conditional on* X

$$\Pr(\pi_z^{00}(X), \pi_z^{01}(X), \pi_z^{11}(X) \geq 0) = 1$$
$$\Pr(\pi_z^{00}(X) + \pi_z^{01}(X) + \pi_z^{11}(X) = 1) = 1$$

$$\mathcal{P}_z(Y(0) = y_0, Y(1) = y_1 \mid X, S_z) = \frac{\pi_z^{y_0 y_1}(X) f_z^{y_0 y_1}(S_z \mid X)}{\pi_z^{00}(X) f_z^{00}(S_z \mid X) + \pi_z^{01}(X) f_z^{01}(S_z \mid X) + \pi_z^{11}(X) f_z^{11}(S_z \mid X)}$$

In Assumption 3′, judge z's "experiment" $f_z^{00}, f_z^{01}, f_z^{11}$ in the sense of Blackwell (1953) remains the same as in Assumption 3.[14] However, her beliefs about the conditional distribution of the unobserved state $\pi_z^{00}, \pi_z^{01}, \pi_z^{11}$ are left completely unspecified. The following lemma establishes sufficient conditions under which Theorem 1 still holds when we replace Assumption 3 by Assumption 3′.

Lemma 1. *Suppose Assumptions 1, 2 and 3′ hold, and let* $x \in \text{supp}(X)$*. If we additionally assume that for all* $z \in \{0, \ldots, k\}$*, either*

1. *Beliefs about the relative proportion of always-appearers and never-appearers are accurate, so*

$$\frac{\pi_z^{11}(x)}{\pi_z^{00}(x) + \pi_z^{11}(x)} = \frac{\pi^{11}(x)}{\pi^{00}(x) + \pi^{11}(x)}$$

   or

2. *Signals are only informative about treatment effects, so*

$$f_z^{00}(s \mid x) = f_z^{11}(s \mid x)$$

   for all $s \in \text{supp}(S_z \mid X = x)$

then the identified set for conditional average counterfactuals $\mathcal{Y}(x)$ *is the same as in Theorem 1.*

If neither of the above conditions hold, judges may misperceive the distribution of their signal conditional on $X = x$ and the event $Y(1) - Y(0) = 0$. This could cause them to make systematic mistakes when ranking defendants based on inaccurate predictions of treatment effects $\mathcal{E}_z[Y(1) - Y(0) \mid X = x, S_z]$, potentially violating positive selection on gains and by extension Theorem 1. However, when either of the above conditions hold, judges will rank defendants correctly, but they will act as if they have a higher (lower) cost of supervision $\lambda_z(x)$ if their perceived CATE $\pi_z^{01}(x)$ is smaller (larger) than the true CATE $\pi^{01}(x)$. So, under the assumptions of Lemma 1, it is impossible to distinguish preferences from beliefs. For example, a judge with a moderate cost of supervision who believes that supervision is moderately

[14] In the literature on robust predictions, this is referred to as her information structure.

effective might behave in the same way as a judge with a high cost of supervision who believes that supervision is highly effective. While this invalidates the bounds on preferences in Corollary 1, it does not affect the results from Theorem 1 since there must still be positive selection on gains regardless of judges' preferences.

Lemma 1 is only a useful generalization of Theorem 1 if we can justify the sufficient conditions it imposes. To this end, models of rational inattention are an important special case to consider (Sims, 2003). In these models, signals are endogenously determined and information is costly to obtain. Consequently, it cannot be optimal for judges to incur the cost of acquiring information to distinguish between states of the world that are not payoff-relevant. Given the preferences specified in Assumption 3, $Y(1) - Y(0)$ represents judges' payoff-relevant unobserved state. So, the condition that signals are only informative about treatment effects will arise endogenously in a large class of rational inattention models.

4.4 Incorporating Taste Shocks

According to Assumption 3, judges make decisions that are not only based on accurate beliefs, but also contain no noise. Any noise contained in judge z's signal is integrated out when she forms her posterior. As a result, putting aside the possibility of indifference, if judge z makes a different decision for two observably identical defendants, she must have a received a signal that the one she chose to supervise had predictably higher returns from supervision. This contradicts a growing literature documenting how seemingly irrelevant factors may influence judges' decisions, such as the defendant's facial features (Ludwig and Mullainathan, 2024), or whether the local football team recently lost a game (Eren and Mocan, 2018). To allow for noisy decisions, I augment Assumption 3 by including taste shocks $\varepsilon_z(0), \varepsilon_z(1)$ that judges receive at arraignment.

Assumption 3′ (Assumption 3 with Taste Shocks)**.** *For all* $z \in \{0, \dots, k\}$,

$$\begin{aligned} Y(0), Y(1) &\perp\!\!\!\perp D(z) \mid X, S_z, \varepsilon_z(0), \varepsilon_z(1) \\ D(z) &\in \arg\max_{d \in \{0,1\}} \mathbb{E}[\mathcal{U}_z(d) \mid X, S_z, \varepsilon_z(0), \varepsilon_z(1)] \\ \mathcal{U}_z(d) &= Y(d) - \lambda_z(X)d + \varepsilon_z(d) \end{aligned}$$

For defendants with $X = x$, judges assign supervision if their *ex ante* expected benefit of supervision exceeds their perceived cost of supervision

$$\mathbb{E}[Y(1) - Y(0) \mid X = x, S_z] > \lambda_z(x) + \varepsilon_z(0) - \varepsilon_z(1)$$

where the difference in taste shocks $\varepsilon_z(0) - \varepsilon_z(1)$ can be interpreted as an unobserved cost. Under this model, judges might exhibit negative selection on gains if there is a positive correlation between their unobserved costs and *ex ante* expected benefits. However, to rule out negative selection on gains, it suffices to assume that taste shocks are idiosyncratic. The following lemma establishes that in this case, Theorem 1 still holds when we replace Assumption 3 by Assumption 3′.

Lemma 2. *Suppose Assumptions 1, 2 and 3′ hold, and let* $x \in \text{supp}(X)$*. If we additionally assume that taste shocks are idiosyncratic*

$$\varepsilon_z(0), \varepsilon_z(1) \perp\!\!\!\perp Y(0), Y(1), S_z \mid X = x$$

then the identified set for conditional average counterfactuals $\mathcal{Y}(x)$ *is the same as in Theorem 1.*

Unlike Lemma 1, under the assumptions of Lemma 2 judges may not correctly rank defendants based on predicted treatment effects $\mathbb{E}[Y(1) - Y(0) \mid X = x, S_z]$ when deciding who to supervise. These misrankings are driven by unobserved costs that are assumed to be idiosyncratic, but whose distribution is unspecified and may be heterogeneous across judges. Interestingly, this may cause violations of a stronger notion of positive selection on gains that requires judge-specific MTE curves $\text{MTE}_z(u \mid x)$ to be decreasing, as they are under Assumption 3.[15] Nevertheless, the weaker condition $\text{ATU}_z(x) \leq \text{ATT}_z(x)$ still holds, and therefore so does Theorem 1. By contrast, Corollary 1 no longer holds under the assumptions of Lemma 2. Even if we assume that unobserved costs are mean zero, we cannot distinguish judges with weak signals and small unobserved costs from judges with strong signals and large unobserved costs. Intuitively, the problem is that preferences $\lambda_z(x)$ are defined relative to the scale of the distribution of unobserved costs.

5 Empirical Results under the Behavioral Model

5.1 Estimation Strategy

The sharp identified set for conditional average counterfactuals $\mathcal{Y}(x)$ defined in Theorem 1 has the advantage of admitting a simple expression in terms of population moments and being robust to the presence of certain unobserved costs and inaccurate beliefs. However, performing estimation and inference on $\mathcal{Y}(x)$ requires overcoming several challenges. First, since X and Z are high-dimensional, it will typically not be possible to estimate objects like $\mathbb{E}[Y \mid X, Z]$ at a $n^{-1/2}$ rate in the L^∞ sense. Second, even if we could come up with a $n^{-1/2}$-consistent estimator of $\mathbb{E}[Y \mid X = x, Z = z]$, we still cannot construct a locally asymptotically unbiased estimator of, for example, $\max_z\{\mathbb{E}[Y \mid X = x, Z = z]\}$ (Hirano and Porter, 2012). To make progress, we will estimate treatment effect parameters of the form

$$\theta_0 = \mathbb{E}[Y(1) - Y(0) \mid X \in \mathcal{X}]$$

where the event $X \in \mathcal{X}$ occurs with reasonably high probability, for the sake of precision. The sharp identified set for θ_0 is simply

$$\begin{aligned} &\underline{\theta}_0 \leq \theta_0 \leq \overline{\theta}_0 \\ &\underline{\theta}_0 := \mathbb{E}[\underline{\tau}(X) \mid X \in \mathcal{X}] \\ &\overline{\theta}_0 := \mathbb{E}[\overline{\tau}(X) \mid X \in \mathcal{X}] \end{aligned}$$

where $\underline{\tau}(X)$ and $\overline{\tau}(X)$ are defined as in Theorem 1. This follows from the fact that the partially identified sets in Theorem 1 are rectangular over X.

The technique I employ to estimate bounds on θ_0 comes from Semenova (2024), who takes advantage of the fact that the expectation operator $\mathbb{E}[\cdot \mid X \in \mathcal{X}]$ smooths over the non-differentiability of the $\min\{\cdot\}$ and $\max\{\cdot\}$ functions contained in $\underline{\tau}(X)$ and $\overline{\tau}(X)$. This allows for $n^{-1/2}$ consistent, joint asymptotically normal estimation of the bounds on θ_0 so long as we can estimate objects like $\mathbb{E}[Y \mid X, Z]$ at a $n^{-1/4}$ rate in the L^∞ sense and the margin assumption from Mammen and Tsybakov (1999) holds. Intuitively, we need to be able to approximate conditional expectation functions sufficiently well (say, with a machine learning algorithm), and there must not be too high a concentration of judges who are near maximizers or

[15] Judge-specific MTE curves are guaranteed to be decreasing under the additional restriction that taste shocks admit a log-concave density function.

minimizers of the expressions in $\underline{\tau}(X)$ and $\overline{\tau}(X)$. Under these conditions, Semenova (2024) shows that we may directly apply double machine learning techniques from Chernozhukov et al. (2018) to efficiently estimate bounds on θ_0, despite the fact that the score functions are only directionally differentiable.

For example, consider letting $\mathcal{X}$ be the support of X and estimating the sharp lower bound on the average treatment effect given by

$$\underline{\theta}_0 = \mathbb{E}\left[\max_z \left\{\frac{\max_{\tilde{z}}\{\mathbb{E}[Y \mid X, Z = \tilde{z}]\} - \mathbb{E}[Y \mid X, Z = z]}{1 - \mathbb{E}[D \mid X, Z = z]}\right\}\right]$$

assuming $\mathbb{E}[D \mid X, Z = z] \neq 1$ with probability 1. Let $z^*(X)$ and $\tilde{z}^*(X)$ denote the maximizers of the outer and inner maximands, respectively, which I assume are unique with probability 1 for the sake of simplicity. Also, define

$$\begin{aligned} f_0(x, z) &:= \mathbb{E}[Y \mid X = x, Z = z] \\ g_0(x, z) &:= \mathbb{E}[D \mid X = x, Z = z] \\ h_0(x, z) &:= \Pr(Z = z \mid X = x) \\ W &:= (Y, D, Z, X) \end{aligned}$$

I estimate $\underline{\theta}_0$ using the following Neyman orthogonal score

$$\begin{aligned} \psi(W; \underline{\theta}, f, g, h) = &\frac{f(X, \tilde{z}^*(X)) - f(X, z^*(X))}{1 - g(X, z^*(X))} - \underline{\theta} \\ &- \frac{\mathbb{1}\{Z = z^*(X)\}}{h(X, z^*(X))} \frac{1}{1 - g(X, z^*(X))}(Y - f(X, z^*(X))) \\ &+ \frac{\mathbb{1}\{Z = z^*(X)\}}{h(X, z^*(X))} \frac{f(X, \tilde{z}^*(X)) - f(X, z^*(X))}{(1 - g(X, z^*(X)))^2}(D - g(X, z^*(X))) \\ &+ \frac{\mathbb{1}\{Z = \tilde{z}^*(X)\}}{h(X, \tilde{z}^*(X))} \frac{1}{1 - g(X, z^*(X))}(Y - f(X, \tilde{z}^*(X))) \end{aligned}$$

For ease of notation, I suppress the dependence of $z^*(X)$ and $\tilde{z}^*(X)$ on f and g. are By construction, $\underline{\theta}_0$ satisfies the moment condition $\mathbb{E}[\psi(W; \underline{\theta}_0, f_0, g_0, h_0)] = 0$. Given an i.i.d. sample $\{W_i\}_{i=1}^n$, I estimate $\underline{\theta}_0$ by $\hat{\underline{\theta}}$ satisfying the empirical analog

$$\frac{1}{n}\sum_{i=1}^{n} \psi(W_i; \hat{\underline{\theta}}, \hat{f}, \hat{g}, \hat{h}) = 0$$

where $\hat{f}$, $\hat{g}$, and $\hat{h}$ are cross-validated gradient-boosted trees models. Importantly, the predictions $\hat{f}(X_i, z)$, $\hat{g}(X_i, z)$, and $\hat{h}(X_i, z)$ used to construct ψ are out-of-sample predictions formed during 30-fold cross-fitting, which occurs after cross-validation.[16] I estimate the asymptotic variance of $\hat{\underline{\theta}}$ by

$$\hat{\sigma}^2 := \frac{1}{n}\sum_{i=1}^{n} \psi(W_i; \hat{\underline{\theta}}, \hat{f}, \hat{g}, \hat{h})^2$$

Finally, I repeat this process 30 times to reduce noise arising from the sample splitting required to construct out-of-sample predictions, aggregating the results as recommended by Chernozhukov et al. (2018).

[16] For additional details on cross-fitting, see Chernozhukov et al. (2018).

5.2 Preliminary Results

Table 2 presents estimated bounds on average counterfactuals, the ATE, ATT, and ATU.

Table 2: Estimated Bounds on Treatment Effect Parameters

	Full Sample	2022 Onwards
ATE	(0.029, 0.099)	(0.042 0.115)
	[0.007, 0.110]	[0.008 0.131]
ATT	(0.325, 0.486)	(0.426 0.477)
	[0.123, 0.594]	[0.149 0.631]
ATU	(0.000, 0.059)	(0.000 0.072)
	[0.000, 0.064]	[0.000 0.080]
$\mathbb{E}[Y(0)]$	(0.877, 0.892)	(0.854 0.859)
	[0.867, 0.910]	[0.837 0.888]
$\mathbb{E}[Y(1)]$	(0.921, 0.976)	(0.901 0.968)
	[0.912, 0.980]	[0.887 0.975]
Supervision Rate	0.092	0.104
Court Appearance Rate	0.922	0.903
Number of Judges	167	139
Number of Cases	75,262	43,994

Notes: Bounds on parameters are derived from Theorem 1. Double machine learning estimated bounds are reported in parentheses. 90% confidence intervals for parameters based on Stoye (2009) are reported below in square brackets.

The average deterrence effect of supervised release among defendants charged with non-qualifying offenses is economically significant, increasing the court appearance rate (or equivalently, decreasing the failure to appear rate) by between 2.9% and 9.9%; it is also statistically significant at the 5% level. Effects are somewhat larger in the latter half of the sampling period in which court appearance rates were lower and supervision rates were higher. There is strong positive selection on gains, with the lower confidence bound on the ATT (12.3%) being nearly twice as large as the upper confidence bound on the ATU (6.4%). Bounds on $\mathbb{E}[Y(0)]$ imply that if the supervised release program did not exist and all defendants with non-qualifying offenses were released on recognizance, the court appearance rate would have been between 2.9% and 4.4% lower over the entire sampling period.

Given the limitations of my current dataset (in the form of sample restrictions and imperfect conditioning variables), I consider these results to be a promising proof of concept. In future versions of this paper, I hope to apply the same methodology to a more complete dataset from the Office of Court Administration.

6 Appendix Tables

Table 3: Test of Quasi-Random Judge Assignment

	TSLS SR Fitted Values	IJIVE SR Fitted Values
Representation Type		
Public Defender	-0.00051^{*}	-0.00042^{*}
	(0.00028)	(0.00023)
Retained Attorney	-0.00004	-0.00003
	(0.00043)	(0.00036)
Assigned Counsel (18B)	0.00046	0.00031
	(0.00049)	(0.00041)
No Lawyer at Arraignment	0.00057	0.00049
	(0.00080)	(0.00067)
Charge Severity		
Unclassified Misdemeanor	0.00029	0.00013
	(0.00161)	(0.00134)
Class B Misdemeanor	-0.00011	-0.00009
	(0.00121)	(0.00101)
Class E Felony	0.00006	-0.00027
	(0.00090)	(0.00075)
Class D Felony	-0.00049	-0.00066
	(0.00093)	(0.00078)
Class C Felony	0.00373^{*}	0.00274
	(0.00219)	(0.00181)
Class B Felony	0.00084	0.00043
	(0.00106)	(0.00088)
Class A Felony	-0.00011	-0.00064
	(0.00273)	(0.00230)
Charge Category		
Attempt	-0.00004	0.00001
	(0.00135)	(0.00113)
Property	-0.00128^{***}	-0.00102^{***}
	(0.00045)	(0.00038)
Larceny	-0.00006	-0.00003
	(0.00061)	(0.00051)
Drug	-0.00108	-0.00084
	(0.00075)	(0.00062)
Aggravated Harassment	-0.00112^{*}	-0.00094^{*}
	(0.00064)	(0.00053)
Endangering Welfare	-0.00032	-0.00036
	(0.00092)	(0.00077)
Robbery	-0.00088	-0.00095
	(0.00189)	(0.00157)
Burglary	0.00258	0.00209
	(0.00202)	(0.00168)
Criminal Trespass	0.00218	0.00180
	(0.00177)	(0.00148)
Criminal Possession of a Weapon	-0.00001	-0.00014
	(0.00116)	(0.00098)
Obstruction	0.00075	0.00072
	(0.00147)	(0.00123)
Other Penal Law	-0.00084	-0.00063
	(0.00070)	(0.00058)
Driving While Intoxicated	-0.00101	-0.00066
	(0.00159)	(0.00132)
Unlicensed Operation of a Vehicle	0.00057	0.00068
	(0.00177)	(0.00147)
Other Vehicle and Traffic Law	0.00124	0.00097
	(0.00184)	(0.00153)
Criminal History		

Table 3: Test of Quasi-Random Judge Assignment

	TSLS SR Fitted Values	IJIVE SR Fitted Values
Felony Count	0.00003	−0.00003
	(0.00017)	(0.00014)
Misdemeanor Count	0.00013*	0.00005
	(0.00007)	(0.00006)
Pending Felony	−0.00062	−0.00074
	(0.00125)	(0.00104)
Pending Misdemeanor	0.00213*	0.00154
	(0.00122)	(0.00102)
Probation or Parole Supervision	−0.00020	−0.00023
	(0.00069)	(0.00058)
Criminal History Unknown	−0.00202	−0.00166
	(0.00267)	(0.00224)
Demographics		
Age Divided by 100	−0.00081	−0.00052
	(0.00109)	(0.00091)
Female	0.00026	0.00027
	(0.00029)	(0.00025)
Hispanic	−0.00026	−0.00022
	(0.00032)	(0.00027)
White	−0.00051	−0.00040
	(0.00032)	(0.00027)
Asian	−0.00106	−0.00088
	(0.00066)	(0.00055)
Court by Year Month FEs	Yes	Yes
Within R^2	0.00073	0.00063
Outcome Mean	0.09228	0.09228
Outcome SD	0.05390	0.05077
P-value of Joint F-test	0.04497	0.17221
Number of Cases	75,262	75,262

This table reports estimated coefficients on defendant characteristics C from OLS regressions of first-stage fitted values on C and court by year month fixed effects T. Each column corresponds to a different procedure used to construct the first-stage fitted values. The first column uses the fitted values from an OLS regression of supervised release D on a full set of judge indicators and T. The second column uses the fitted values from an OLS regression of D on the IJIVE constructed instrument and T. Neither of the first-stage specifications includes C, so if $Z \perp\!\!\!\perp C \mid T$ then the population counterparts of the reported coefficients should be equal to zero. The p-values of F-tests of this null hypothesis are reported at the bottom of the table. Note that the p-value in the second column would be identical if we had used the IJIVE constructed instrument itself as the outcome variable, rather than the corresponding fitted values. Stars denote $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

7 Appendix

7.1 Proof of Theorem 1

Proof. For ease of notation, implicitly condition on the event $X = x$ throughout. First, we will show that under Assumptions 1, 2 and 3, $\mathbb{E}[Y(0)], \mathbb{E}[Y(1)]$ must satisfy Inequalities 6, 7, 8 and 9 for all $z \in \{0, \dots, k\}$. Let $z \in \{0, \dots, k\}$. Then we have

$$\begin{aligned}
\mathbb{E}[Y(0)] &= \mathbb{E}[D(z)Y(0)] + \mathbb{E}[(1 - D(z))Y(0)] \\
&\geq \mathbb{E}[(1 - D(z))Y(0)] \\
&= \mathbb{E}[(1 - D)Y \mid Z = z] && \text{by Assumption 1} \\
\mathbb{E}[Y(1)] &= \mathbb{E}[D(z)Y(1)] + \mathbb{E}[(1 - D(z))Y(1)] \\
&\geq \mathbb{E}[D(z)Y(1)] + \mathbb{E}[(1 - D(z))Y(0)] && \text{by Assumption 2} \\
&= \mathbb{E}[DY \mid Z = z] + \mathbb{E}[(1 - D)Y \mid Z = z] && \text{by Assumption 1} \\
&= \mathbb{E}[Y \mid Z = z] \\
\mathbb{E}[Y(1)] &= \mathbb{E}[D(z)Y(1)] + \mathbb{E}[(1 - D(z))Y(1)] \\
&\leq \mathbb{E}[D(z)Y(1)] + \mathbb{E}[(1 - D(z))] \\
&= \mathbb{E}[DY \mid Z = z] + (1 - \mathbb{E}[D \mid Z = z]) && \text{by Assumption 1}
\end{aligned}$$

This establishes Inequalities 6, 8 and 9. Establishing Inequality 7 requires applying Assumption 3. Let $d \in \{0, 1\}$. Then we have

$$\begin{aligned}
\mathbb{E}[\mathcal{U}_z(d) \mid D(z) = d] &= \mathbb{E}[\mathbb{E}[\mathcal{U}_z(d) \mid S_z, D(z) = d] \mid D(z) = d] \\
&= \mathbb{E}[\mathbb{E}[\mathcal{U}_z(d) \mid S_z] \mid D(z) = d] && \text{by Equation (1) of Assumption 3} \\
&\geq \mathbb{E}[\mathbb{E}[\mathcal{U}_z(1 - d) \mid S_z] \mid D(z) = d] && \text{by Equation (2) of Assumption 3} \\
&= \mathbb{E}[\mathcal{U}_z(1 - d) \mid D(z) = d]
\end{aligned}$$

From this we arrive at the revealed preference inequalities, or obedience conditions, that summarize the empirical content of Assumption 3. In particular,

$$\begin{aligned}
\mathbb{E}[Y(0)] &= \mathbb{E}[\mathcal{U}_z(0)] && \text{by Equation (3) of Assumption 3} \\
&= \mathbb{E}[D(z)]\mathbb{E}[\mathcal{U}_z(0) \mid D(z) = 1] + (1 - \mathbb{E}[D(z)])\mathbb{E}[\mathcal{U}_z(0) \mid D(z) = 0] \\
&\leq \mathbb{E}[D(z)]\mathbb{E}[\mathcal{U}_z(1) \mid D(z) = 1] + (1 - \mathbb{E}[D(z)])\mathbb{E}[\mathcal{U}_z(0) \mid D(z) = 0] \\
&= \mathbb{E}[\mathcal{U}_z(D(z))] \\
&= \mathbb{E}[Y(D(z)) - \lambda_z(x)D(z)] \\
&= \mathbb{E}[Y \mid Z = z] - \lambda_z(x)\mathbb{E}[D \mid Z = z] && \text{by Assumption 1}
\end{aligned}$$

and similarly

$$\begin{aligned}
\mathbb{E}[Y(1)] - \lambda_z(x) &= \mathbb{E}[\mathcal{U}_z(1)] && \text{by Equation (3) of Assumption 3} \\
&= \mathbb{E}[D(z)]\mathbb{E}[\mathcal{U}_z(1) \mid D(z) = 1] + (1 - \mathbb{E}[D(z)])\mathbb{E}[\mathcal{U}_z(1) \mid D(z) = 0] \\
&\leq \mathbb{E}[D(z)]\mathbb{E}[\mathcal{U}_z(1) \mid D(z) = 1] + (1 - \mathbb{E}[D(z)])\mathbb{E}[\mathcal{U}_z(0) \mid D(z) = 0] \\
&= \mathbb{E}[Y \mid Z = z] - \lambda_z(x)\mathbb{E}[D \mid Z = z] && \text{by Assumption 1}
\end{aligned}$$

These correspond to Inequalities 4 and 5 in the main text. Multiply Inequality 4 by $(1 - \mathbb{E}[D \mid Z = z])$ and Inequality 5 by $\mathbb{E}[D \mid Z = z]$ to get

$$\begin{aligned}(1 - \mathbb{E}[D \mid Z = z])\mathbb{E}[Y(0)] &\leq (1 - \mathbb{E}[D \mid Z = z])\mathbb{E}[Y \mid Z = z] \\ &\quad - \lambda_z(x)\mathbb{E}[D \mid Z = z](1 - \mathbb{E}[D \mid Z = z]) \\ \mathbb{E}[D \mid Z = z]\mathbb{E}[Y(1)] &\leq \mathbb{E}[D \mid Z = z]\mathbb{E}[Y \mid Z = z] \\ &\quad + \lambda_z(x)\mathbb{E}[D \mid Z = z](1 - \mathbb{E}[D \mid Z = z])\end{aligned}$$

Adding the resulting inequalities together establishes Inequality 7

$$\mathbb{E}[Y(0)] \leq \mathbb{E}[Y \mid Z = z] - \mathbb{E}[D \mid Z = z]\mathbb{E}[Y(1) - Y(0)]$$

This completes the first part of the proof: since our choice of z was arbitrary, $\mathbb{E}[Y(0)], \mathbb{E}[Y(1)]$ must lie in the identified set $\mathcal{Y}(x)$. It remains to show that the identified set is sharp.

For any $y_0, y_1 \in \mathcal{Y}(x)$, we will construct a joint distribution of counterfactual outcomes, counterfactual treatments, and judge assignment $\tilde{Y}(0), \tilde{Y}(1), \tilde{D}(0), \ldots, \tilde{D}(k), \tilde{Z}$ along with signals $\{\tilde{S}_z\}_{z=0}^k$ and preference parameters $\{\tilde{\lambda}_z(x)\}_{z=0}^k$ such that

1. Conditional average counterfactuals $\mathbb{E}[\tilde{Y}(0)], \mathbb{E}[\tilde{Y}(1)]$ are given by y_0, y_1
2. Assumptions 1, 2 and 3 hold
3. The resulting distribution $\tilde{Y}, \tilde{D}, \tilde{Z}$ matches the distribution of observables Y, D, Z

Let $\tilde{Z} \stackrel{d}{=} Z$ with

$$\tilde{Y}(0), \tilde{Y}(1), \tilde{D}(0), \ldots, \tilde{D}(k) \perp\!\!\!\perp \tilde{Z}$$

so Assumption 1 holds. Without loss of generality, let $\tilde{D}(0), \ldots, \tilde{D}(k)$ be mutually independent given $\tilde{Y}(0), \tilde{Y}(1)$. For all $z \in \{0, \ldots, k\}$, let the joint distribution of $\tilde{Y}(0), \tilde{Y}(1), \tilde{D}(z)$ be given by

$$\begin{aligned}\Pr(\tilde{Y}(0) = 0, \tilde{Y}(1) = 0, \tilde{D}(z) = 0) &= 1 - \mathbb{E}[D(1 - Y) \mid Z = z] - y_1 \\ \Pr(\tilde{Y}(0) = 0, \tilde{Y}(1) = 0, \tilde{D}(z) = 1) &= \mathbb{E}[D(1 - Y) \mid Z = z] \\ \Pr(\tilde{Y}(0) = 0, \tilde{Y}(1) = 1, \tilde{D}(z) = 0) &= y_1 - \mathbb{E}[Y \mid Z = z] \\ \Pr(\tilde{Y}(0) = 0, \tilde{Y}(1) = 1, \tilde{D}(z) = 1) &= \mathbb{E}[Y \mid Z = z] - y_0 \\ \Pr(\tilde{Y}(0) = 1, \tilde{Y}(1) = 0, \tilde{D}(z) = 0) &= 0 \\ \Pr(\tilde{Y}(0) = 1, \tilde{Y}(1) = 0, \tilde{D}(z) = 1) &= 0 \\ \Pr(\tilde{Y}(0) = 1, \tilde{Y}(1) = 1, \tilde{D}(z) = 0) &= \mathbb{E}[(1 - D)Y \mid Z = z] \\ \Pr(\tilde{Y}(0) = 1, \tilde{Y}(1) = 1, \tilde{D}(z) = 1) &= y_0 - \mathbb{E}[(1 - D)Y \mid Z = z]\end{aligned}$$

so Assumption 2 holds, conditional average counterfactuals are given by y_0, y_1, and $\tilde{Y}, \tilde{D}, \tilde{Z} \stackrel{d}{=} Y, D, Z$. Since $y_0, y_1 \in \mathcal{Y}(x)$ by assumption, all of the above probabilities are in $[0, 1]$. All that remains is to show that Assumption 3 holds. For all $z \in \{0, \ldots, k\}$, let $\tilde{S}_z = \tilde{D}(z)$ and let

$$\tilde{\lambda}_z(x) \in \begin{cases} [y_1 - y_0, \infty] & \text{if} \quad \mathbb{E}[D \mid Z = z] = 0 \\ \left[\frac{y_1 - \mathbb{E}[Y|Z=z]}{1 - \mathbb{E}[D|Z=z]}, \frac{\mathbb{E}[Y|Z=z] - y_0}{\mathbb{E}[D|Z=z]}\right] & \text{if} \quad \mathbb{E}[D \mid Z = z] \in (0, 1) \\ [-\infty, y_1 - y_0] & \text{if} \quad \mathbb{E}[D \mid Z = z] = 1 \end{cases}$$

where the above intervals are non-empty since $y_0, y_1 \in \mathcal{Y}(x)$ by assumption so Inequality 7 must hold. Equation (1) of Assumption 3 must hold since $\tilde{D}(z)$ is a degenerate random variable conditional on $\tilde{S}_z$. Equation (2) of Assumption 3 must hold with probability one since

$$\begin{aligned}
&\Pr(\tilde{D}(z) \in \arg\max_{d \in \{0,1\}} \mathbb{E}[\tilde{\mathcal{U}}_z(d) \mid \tilde{S}_z]) \\
&= \Pr(\tilde{D}(z) = 0) \Pr(\tilde{D}(z) \in \arg\max_{d \in \{0,1\}} \mathbb{E}[\tilde{\mathcal{U}}_z(d) \mid \tilde{S}_z] \mid \tilde{D}(z) = 0) \\
&+ \Pr(\tilde{D}(z) = 1) \Pr(\tilde{D}(z) \in \arg\max_{d \in \{0,1\}} \mathbb{E}[\tilde{\mathcal{U}}_z(d) \mid \tilde{S}_z] \mid \tilde{D}(z) = 1) \\
&= (1 - \mathbb{E}[D \mid Z = z]) \mathbb{1}\{\mathbb{E}[\tilde{\mathcal{U}}_z(0) \mid \tilde{D}(z) = 0] \geq \mathbb{E}[\tilde{\mathcal{U}}_z(1) \mid \tilde{D}(z) = 0]\} \\
&+ \mathbb{E}[D \mid Z = z] \mathbb{1}\{\mathbb{E}[\tilde{\mathcal{U}}_z(1) \mid \tilde{D}(z) = 1] \geq \mathbb{E}[\tilde{\mathcal{U}}_z(0) \mid \tilde{D}(z) = 1]\}
\end{aligned}$$

If $\mathbb{E}[D \mid Z = z] = 0$, the first of the above indicators is equal to one since $\Pr(\tilde{D}(z) = 0) = 1$ and

$$\begin{aligned}
&\mathbb{E}[\tilde{\mathcal{U}}_z(0)] \geq \mathbb{E}[\tilde{\mathcal{U}}_z(1)] \\
\iff &\mathbb{E}[\tilde{Y}(0)] \geq \mathbb{E}[\tilde{Y}(1) - \tilde{\lambda}_z(x)] && \text{by Equation (3) of Assumption 3} \\
\iff &\tilde{\lambda}_z(x) \geq y_1 - y_0
\end{aligned}$$

which holds by definition of $\tilde{\lambda}_z(x)$. If $\mathbb{E}[D \mid Z = z] = 1$, the second of the above indicators is equal to one since $\Pr(\tilde{D}(z) = 1) = 1$ and

$$\begin{aligned}
&\mathbb{E}[\tilde{\mathcal{U}}_z(0)] \leq \mathbb{E}[\tilde{\mathcal{U}}_z(1)] \\
\iff &\mathbb{E}[\tilde{Y}(0)] \leq \mathbb{E}[\tilde{Y}(1) - \tilde{\lambda}_z(x)] && \text{by Equation (3) of Assumption 3} \\
\iff &\tilde{\lambda}_z(x) \leq y_1 - y_0
\end{aligned}$$

which holds by definition of $\tilde{\lambda}_z(x)$. Finally, if $\mathbb{E}[D \mid Z = z] \in (0, 1)$, both of the above indicators are equal to one since

$$\begin{aligned}
&\mathbb{E}[\tilde{\mathcal{U}}_z(0) \mid \tilde{D}(z) = 0] \geq \mathbb{E}[\tilde{\mathcal{U}}_z(1) \mid \tilde{D}(z) = 0] \\
&\iff \tilde{\lambda}_z(x) \geq \mathbb{E}[\tilde{Y}(1) - \tilde{Y}(0) \mid \tilde{D}(z) = 0] && \text{by Equation (3) of Assumption 3} \\
&\iff \tilde{\lambda}_z(x) \geq \frac{y_1 - \mathbb{E}[Y \mid Z = z]}{1 - \mathbb{E}[D \mid Z = z]}
\end{aligned}$$

and similarly

$$\begin{aligned}
&\mathbb{E}[\tilde{\mathcal{U}}_z(0) \mid \tilde{D}(z) = 1] \leq \mathbb{E}[\tilde{\mathcal{U}}_z(1) \mid \tilde{D}(z) = 1] \\
&\iff \tilde{\lambda}_z(x) \leq \mathbb{E}[\tilde{Y}(1) - \tilde{Y}(0) \mid \tilde{D}(z) = 1] && \text{by Equation (3) of Assumption 3} \\
&\iff \tilde{\lambda}_z(x) \leq \frac{\mathbb{E}[Y \mid Z = z] - y_0}{\mathbb{E}[D \mid Z = z]}
\end{aligned}$$

both of which hold by definition of $\tilde{\lambda}_z(x)$.

In summary, we have shown that under Assumptions 1, 2 and 3, the sharp identified set for conditional average counterfactuals $\mathbb{E}[Y(0)], \mathbb{E}[Y(1)]$ is given by $\mathcal{Y}(x)$. This set is defined as an intersection of convex polygons on $[0, 1] \times [0, 1]$, and is therefore also a convex polygon on $[0, 1] \times [0, 1]$.

The final part of the proof involves showing that the proposed bounds on conditional average treatment effects correspond to the optimized objectives of the corresponding linear programs. This is a bit tedious, so I have omitted it from the current draft. □

7.2 Proof of Corollary 1

Proof. Almost all of the work to establish Corollary 1 was done in the proof of Theorem 1. First, for any pair $y_0, y_1 \in \mathcal{Y}(x)$, consider the sharp identified set for $\lambda_z(x)$ under Assumptions 1, 2 and 3, and the additional assumption that $\mathbb{E}[Y(0) \mid X = x], \mathbb{E}[Y(1) \mid X = x] = y_0, y_1$. From the first part of the proof of Theorem 1, we know that under these assumptions Inequalities 4 and 5 must hold

$$\begin{aligned} y_0 &\leq \mathbb{E}[Y \mid X = x, Z = z] - \lambda_z(x)\mathbb{E}[D \mid X = x, Z = z] \\ y_1 - \lambda_z(x) &\leq \mathbb{E}[Y \mid X = x, Z = z] - \lambda_z(x)\mathbb{E}[D \mid X = x, Z = z] \end{aligned}$$

which can alternatively be expressed as $\lambda_z(x) \in \mathcal{L}_z(x; y_0, y_1)$ where

$$\mathcal{L}_z(x; y_0, y_1) := \begin{cases} [y_1 - y_0, \infty] & \text{if} \quad \mathbb{E}[D \mid X = x, Z = z] = 0 \\ \left[\frac{y_1 - \mathbb{E}[Y|X=x,Z=z]}{1-\mathbb{E}[D|X=x,Z=z]}, \frac{\mathbb{E}[Y|X=x,Z=z]-y_0}{\mathbb{E}[D|X=x,Z=z]}\right] & \text{if} \quad \mathbb{E}[D \mid X = x, Z = z] \in (0,1) \\ [-\infty, y_1 - y_0] & \text{if} \quad \mathbb{E}[D \mid X = x, Z = z] = 1 \end{cases}$$

But then sharpness follows immediately from the fact that this is exactly the same set considered in the second part of the proof of Theorem 1. We arrive at the sharp identified set for $\lambda_z(x)$ under Assumptions 1, 2 and 3 by taking a union of the sets $\mathcal{L}_z(x; y_0, y_1)$ over $\mathcal{Y}(x)$. □

7.3 Proof of Lemma 1

Proof. Notice that the assumptions of Theorem 1 are a special case of the assumptions of Lemma 1 in which beliefs are accurate. So, if we can show that $\mathcal{Y}(x)$ is an identified set for conditional average counterfactuals under the assumptions of Lemma 1, it must be the sharp identified set. Consider the relationship between $\mathcal{E}_z[Y(1) - Y(0) \mid X = x, S_z]$ and $\mathbb{E}[Y(1) - Y(0) \mid X = x, S_z]$. Under Assumption 3′, we have

$$\begin{aligned} \mathcal{E}_z[Y(1) - Y(0) \mid X = x, S_z] &= \mathcal{P}_z(Y(0) = 0, Y(1) = 1 \mid X = x, S_z) \qquad \text{by Assumption 2} \\ &= \frac{\pi_z^{01}(x) f_z^{01}(S_z \mid x)}{\pi_z^{00}(x) f_z^{00}(S_z \mid x) + \pi_z^{01}(x) f_z^{01}(S_z \mid x) + \pi_z^{11}(x) f_z^{11}(S_z \mid x)} \end{aligned}$$

and similarly

$$\mathbb{E}[Y(1) - Y(0) \mid X = x, S_z] = \frac{\pi^{01}(x) f_z^{01}(S_z \mid x)}{\pi^{00}(x) f_z^{00}(S_z \mid x) + \pi^{01}(x) f_z^{01}(S_z \mid x) + \pi^{11}(x) f_z^{11}(S_z \mid x)}$$

Some algebraic manipulation yields

$$\begin{aligned} \mathbb{E}[Y(1) - Y(0) \mid X = x, S_z] &= \frac{1}{1 + R\frac{1-\pi^{01}(x)}{\pi^{01}(x)}\frac{\pi_z^{01}(x)}{1-\pi_z^{01}(x)}\frac{1-\mathcal{E}_z[Y(1)-Y(0)|X=x,S_z]}{\mathcal{E}_z[Y(1)-Y(0)|X=x,S_z]}} \\ R &:= \frac{\frac{\pi^{00}(x) f_z^{00}(S_z|x) + \pi^{11}(x) f_z^{11}(S_z|x)}{\pi^{00}(x) + \pi^{11}(x)}}{\frac{\pi_z^{00}(x) f_z^{00}(S_z|x) + \pi_z^{11}(x) f_z^{11}(S_z|x)}{\pi_z^{00}(x) + \pi_z^{11}(x)}} \end{aligned}$$

The key idea of the proof is that under the additional assumption of Lemma 1, $R = 1$ with probability one. In this case, $\mathbb{E}[Y(1) - Y(0) \mid X = x, S_z]$ is a strictly increasing function of $\mathcal{E}_z[Y(1) - Y(0) \mid X = x, S_z]$ on $[0, 1]$ given by

$$g(y; \pi^{01}(x), \pi_z^{01}(x)) = \frac{1}{1 + \frac{1-\pi^{01}(x)}{\pi^{01}(x)}\frac{\pi_z^{01}(x)}{1-\pi_z^{01}(x)}\frac{1-y}{y}}$$

Conditional on $X = x$, Assumption 3′ holds if and only if $Y(0), Y(1) \perp\!\!\!\perp D(z) \mid X = x, S_z$ and

$$D(z) \begin{cases} = 0 & \text{if} \quad \mathcal{E}_z[Y(1) - Y(0) \mid X = x, S_z] < \lambda_z(x) \\ \in \{0, 1\} & \text{if} \quad \mathcal{E}_z[Y(1) - Y(0) \mid X = x, S_z] = \lambda_z(x) \\ = 1 & \text{if} \quad \mathcal{E}_z[Y(1) - Y(0) \mid X = x, S_z] > \lambda_z(x) \end{cases}$$

But this is equivalent to

$$D(z) \begin{cases} = 0 & \text{if} \quad \mathbb{E}[Y(1) - Y(0) \mid X = x, S_z] < g(\lambda_z(x); \pi^{01}(x), \pi_z^{01}(x)) \\ \in \{0, 1\} & \text{if} \quad \mathbb{E}[Y(1) - Y(0) \mid X = x, S_z] = g(\lambda_z(x); \pi^{01}(x), \pi_z^{01}(x)) \\ = 1 & \text{if} \quad \mathbb{E}[Y(1) - Y(0) \mid X = x, S_z] > g(\lambda_z(x); \pi^{01}(x), \pi_z^{01}(x)) \end{cases}$$

So, under the additional assumption of Lemma 1, Assumption 3′ is equivalent to Assumption 3 with preference parameters given by $\{g(\lambda_z(x); \pi^{01}(x), \pi_z^{01}(x))\}_{z=0}^k$. We may apply the first part of the proof of Theorem 1 to conclude that $\mathcal{Y}(x)$ is an identified set for conditional average counterfactuals. As previously discussed, sharpness follows immediately.

This proof is a bit sloppy, since it does not consider the possibility that $\pi^{01}(x)$ or $\pi_z^{01}(x)$ are equal to zero or one, that $\lambda_z(x)$ is outside of $[0, 1]$, or that the densities $\{f_z^{00}, f_z^{01}, f_z^{11}\}$ are not well-defined. I hope to fix this in a future draft. It is also worth noting that, under the assumptions of Lemma 1, Corollary 1 yields the sharp identified set for $g(\lambda_z(x); \pi^{01}(x), \pi_z^{01}(x))$, rather than $\lambda_z(x)$. It is not possible to distinguish preferences $\lambda_z(x)$ from beliefs $\pi_z^{01}(x)$; in the case of an interior solution, the sharp identified set for $\lambda_z(x)$ is $(0, 1)$. I hope to flesh out this idea, which is closely related to Rambachan (2024), in a future draft. □

7.4 Proof of Lemma 2

Proof. Notice that the assumptions of Theorem 1 are a special case of the assumptions of Lemma 2 in which taste shocks are degenerate and equal to zero. So, if we can show that $\mathcal{Y}(x)$ is an identified set for conditional average counterfactuals under the assumptions of Lemma 2, it must be the sharp identified set. Since Inequalities 6, 8 and 9 must hold regardless of our behavioral model, all that remains is to establish Inequality 7. Let $z \in \{0, \ldots, k\}$. If $\mathbb{E}[D \mid X = x, Z = z] = 0$, then Inequality 7 holds trivially since $\mathbb{E}[Y \mid X = x, Z = z] = \mathbb{E}[Y(0) \mid X = x]$ by Assumption 1. Now, suppose $\mathbb{E}[D \mid X = x, Z = z] > 0$. Then by Assumption 1, we have

$$\text{ATT}_z(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x, D(z) = 1] = \frac{\mathbb{E}[Y \mid X = x, Z = z] - \mathbb{E}[Y(0) \mid X = x]}{\mathbb{E}[D \mid X = x, Z = z]}$$

so Inequality 7 is equivalent to $\text{ATT}_z(x) \geq \text{ATE}(x)$. For ease of notation, define

$$\begin{aligned} \delta_z &:= \varepsilon_z(0) - \varepsilon_z(1) \\ V_z &:= \mathbb{E}[Y(1) - Y(0) \mid X = x, S_z] \end{aligned}$$

Then by Assumption 3″, conditional on $X = x$ we have

$$D(z) \begin{cases} = 0 & \text{if} \quad V_z < \lambda_z(x) + \delta_z \\ \in \{0, 1\} & \text{if} \quad V_z = \lambda_z(x) + \delta_z \\ = 1 & \text{if} \quad V_z > \lambda_z(x) + \delta_z \end{cases}$$

To avoid technical issues arising from tie-breaking, suppose $D(z) = \mathbb{1}\{V_z \geq \lambda_z(x) + \delta_z\}$; this is something I hope to fix in a future draft. Then we have

$$\begin{aligned}
\text{ATT}_z(x) &= \mathbb{E}[Y(1) - Y(0) \mid X = x, D(z) = 1] \\
&= \mathbb{E}[Y(1) - Y(0) \mid X = x, V_z \geq \lambda_z(x) + \delta_z] \\
&= \mathbb{E}[\mathbb{E}[Y(1) - Y(0) \mid X = x, V_z, \delta_z] \mid X = x, V_z \geq \lambda_z(x) + \delta_z] \\
&= \mathbb{E}[\mathbb{E}[Y(1) - Y(0) \mid X = x, V_z] \mid X = x, V_z \geq \lambda_z(x) + \delta_z] \\
&= \mathbb{E}[V_z \mid X = x, V_z \geq \lambda_z(x) + \delta_z]
\end{aligned}$$

In addition, we have $V_z \perp\!\!\!\perp \lambda_z(x) + \delta_z \mid X = x$. Finally, we have

$$\text{ATE}(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x] = \mathbb{E}[V_z \mid X = x]$$

The desired result follows from the fact that for any two independent scalar random variables A and B, $\mathbb{E}[A \mid A \geq B] \geq \mathbb{E}[A]$. □

References

Ackerberg, D. A. and P. J. Devereux (2009). Improved jive estimators for overidentified linear models with and without heteroskedasticity. *The Review of Economics and Statistics 91*(2), 351–362.

Albright, A. (2022). No money bail, no problems? trade-offs in a pretrial automatic release program.

Angelova, V., W. S. Dobbie, and C. Yang (2023). Algorithmic recommendations and human discretion. Technical report, National Bureau of Economic Research.

Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association 91*(434), 444–455.

Arnold, D., W. Dobbie, and P. Hull (2022, September). Measuring racial discrimination in bail decisions. *American Economic Review 112*(9), 2992–3038.

Bergemann, D. and S. Morris (2016). Bayes correlated equilibrium and the comparison of information structures in games. *Theoretical Economics 11*(2), 487–522.

BJS (2023). Jail Inmates in 2022 – Statistical Tables. Technical report, Bureau of Justice Statistics. (accessed June 2024).

Blackwell, D. (1953). Equivalent comparisons of experiments. *The Annals of Mathematical Statistics 24*(2), 265–272.

Blandhol, C., J. Bonney, M. Mogstad, and A. Torgovitsky (2022). When is tsls actually late? Technical report, National Bureau of Economic Research Cambridge, MA.

Canay, I. A., M. Mogstad, and J. Mountjoy (2023). On the use of outcome tests for detecting bias in decision making. *Review of Economic Studies*, rdad082.

CCR (2022). The Civil Rights Implications of Cash Bail. Technical report, U.S. Commission on Civil Rights. (accessed June 2024).

Chan, D. C., M. Gentzkow, and C. Yu (2022, 01). Selection with Variation in Diagnostic Skill: Evidence from Radiologists*. *The Quarterly Journal of Economics 137*(2), 729–783.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018, 01). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal 21*(1), C1–C68.

Chyn, E., B. Frandsen, and E. C. Leslie (2024). Examiner and judge designs in economics: A practitioner's guide. Technical report, National Bureau of Economic Research.

DCJS (2018). New York State Bar Association, 2018 Annual Meeting, Materials for Panel Discussion on Bail Reform. Technical report, Division of Criminal Justice Services.

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

D'Haultfœuille, X. and A. Maurel (2013). Inference on an extended roy model, with an application to schooling decisions in france. *Journal of Econometrics 174*(2), 95–106.

Eren, O. and N. Mocan (2018, July). Emotional judges and unlucky juveniles. *American Economic Journal: Applied Economics 10*(3), 171–205.

Frandsen, B., L. Lefgren, and E. Leslie (2023, January). Judging judge fixed effects. *American Economic Review 113*(1), 253–77.

Gualdani, C. and S. Sinha (2019). Identification in discrete choice models with imperfect information. arXiv preprint arXiv:1911.04529.

Heckman, J. J. and E. J. Vytlacil (1999). Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proceedings of the national Academy of Sciences 96*(8), 4730–4734.

Heckman, J. J. and E. J. Vytlacil (2007). Chapter 71 Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast their Effects in New Environments. In J. J. Heckman and E. E. Leamer (Eds.), *Handbook of Econometrics*, Volume 6, pp. 4875–5143. Elsevier.

Hirano, K. and J. R. Porter (2012). Impossibility results for nondifferentiable functionals. *Econometrica 80*(4), 1769–1790.

Imbens, G. W. and J. D. Angrist (1994). Identification and estimation of local average treatment effects. *Econometrica 62*(2), 467–475.

Katz, M. (2023). NYC's supervised release program swelled after bail reform. Now it may be overwhelmed. (accessed June 2024).

Kleinberg, J., H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan (2017, 08). Human Decisions and Machine Predictions*. *The Quarterly Journal of Economics 133*(1), 237–293.

Kolesár, M. (2013). Estimation in an instrumental variables model with treatment effect heterogeneity. Technical report.

Leslie, E. and N. G. Pope (2017). The unintended impact of pretrial detention on case outcomes: Evidence from new york city arraignments. *The Journal of Law and Economics 60*(3), 529–557.

Ludwig, J. and S. Mullainathan (2024). Machine learning as a tool for hypothesis generation. *The Quarterly Journal of Economics 139*(2), 751–827.

Mammen, E. and A. B. Tsybakov (1999). Smooth discrimination analysis. *The Annals of Statistics 27*(6), 1808–1829.

Manski, C. F. (1990). Nonparametric bounds on treatment effects. *The American Economic Review 80*(2), 319–323.

Manski, C. F. (1997). Monotone treatment response. *Econometrica 65*(6), 1311–1334.

Manski, C. F. and J. V. Pepper (2000). Monotone instrumental variables: With an application to the returns to schooling. *Econometrica 68*(4), 997–1010.

Mayor's Office of Criminal Justice (2024). Program: Supervised Release. https://criminaljustice.cityofnewyork.us/programs/supervised-release/. (accessed June 2024).

MDRC (2020). Pursuing Pretrial Justice Through an Alternative to Bail: Findings from an Evaluation of New York City's Supervised Release Program. Technical report, MDRC. (accessed June 2024).

New York State Senate (2023a). Criminal Procedure (CPL) Chapter 11-A, Part 3, Title P, Article 510. https://www.nysenate.gov/legislation/laws/CPL/510.10. (accessed June 2024).

New York State Senate (2023b). Criminal Procedure (CPL) Chapter 11-A, Part 3, Title P, Article 530. https://www.nysenate.gov/legislation/laws/CPL/530.40. (accessed June 2024).

OCA (2024). New York State Office of Court Administration Pretrial Release Data, NYC CSV Extract from 1/1/20 to 1/1/24. (accessed June 2024).

Rambachan, A. (2024, 05). Identifying Prediction Mistakes in Observational Data*. *The Quarterly Journal of Economics*, qjae013.

Rivera, R. (2023). Release, detain, or surveil?

Roy, A. D. (1951). Some thoughts on the distribution of earnings. *Oxford Economic Papers 3*(2), 135–146.

Semenova, V. (2024). Aggregated intersection bounds and aggregated minimax values.

Sims, C. A. (2003). Implications of rational inattention. *Journal of Monetary Economics 50*(3), 665–690. Swiss National Bank/Study Center Gerzensee Conference on Monetary Policy under Incomplete Information.

Stoye, J. (2009). More on confidence intervals for partially identified parameters. *Econometrica 77*(4), 1299–1315.

Vytlacil, E. (2002). Independence, monotonicity, and latent index models: An equivalence result. *Econometrica 70*(1), 331–341.