Conviction, Incarceration, and Policy Effects in the Criminal Justice System[*]

Vishal Kamat,[†] Samuel Norris,[‡] Matthew Pecenco[§]

August 7, 2023

Abstract

The criminal justice system affects millions of Americans through incarceration as well as conviction and the resulting criminal record. In this paper, we introduce a new method for credibly estimating the effects of both conviction and incarceration using randomly assigned judges as instruments for treatment. Misdemeanor convictions, especially for defendants with a shorter criminal record, cause an increase in the number of new offenses committed over the following five years. Felony incarceration, in contrast, decreases future crime through an expensive incapacitation effect. Our method allows the researcher to decompose these treatment effects into their constituent parts as well as estimate the effect of broader policies; we find that courts could simultaneously increase leniency and reduce crime through policies that target defendants accused of serious misdemeanors.

[*] We thank Marit Rehavi and seminar audiences at Brown, UC Santa Barbara, UC Irvine, UC Davis, Stanford, and SOLE for their feedback. Alison Lodermeier provided outstanding research assistance.

[†] Toulouse School of Economics, vishal.kamat@tse-fr.eu

[‡] University of British Columbia, sam.norris@ubc.ca

[§] Brown University. matthew_pecenco@brown.edu

Decision-makers such as judges and doctors can profoundly affect individual outcomes through the choices they make. As a result, quasi-randomly assigned decision-makers are often used as instruments when estimating the causal effect of treatments as varied as incarceration, hospital quality, and disability insurance (Kling, 2006; Doyle et al., 2015; Maestas, Mullen and Strand, 2013). The majority of prior work has used these *examiner designs* to estimate the effect of a single treatment, such as incarcerating a defendant. However, examiners often make choices over more than two options—for example, whether to acquit, convict, or incarcerate. This multi-dimensionality threatens the validity of the assumptions needed for 2SLS to deliver an interpretable treatment effect (Kirkeboen, Leuven and Mogstad, 2016).

In this paper, we develop a new framework to analyze settings with multiple observed treatments and randomly-assigned examiners. We first study the common approach of instrumenting only for the treatment of interest, and show that the resulting 2SLS estimand is interpretable as a treatment effect only under restrictive assumptions that are unlikely to be met in examiner settings. Other approaches that explicitly accommodate multiple treatments, such as unordered monotonicity (Heckman and Pinto, 2018), also rule out likely substitution patterns. To account for these issues, we study a model of treatment assignment that imposes only weak assumptions on substitution patterns across judges, and which can be translated into a tractable structural form for estimation. Our approach allows the researcher to target traditional IV estimands and their margin-specific subcomponents, as well as a wide variety of policy-relevant treatment effects (Heckman and Vytlacil, 2005).

We use this new framework to analyze the effects of conviction and incarceration in the US criminal justice system. Both of these criminal sanctions are common—17 million new cases are filed each year in state courts alone (Court Statistics Project, 2018)—but the evidence on their effects comes from very different environments. Research on conviction has focused on settings where defendants face minor misdemeanor charges, and so are nearly never incarcerated no matter the case outcome. This allows a clean estimate of the effect of being prosecuted (Agan, Doleac and Harvey, 2022) or convicted (Agan et al., 2023) versus a counterfactual of no sanction. In contrast, work on incarceration has largely used a counterfactual of a more serious felony conviction (Norris, Pecenco and Weaver, 2021; Rose and Shem-Tov, 2021). This has made it difficult to understand why conviction and incarceration appear to have opposite effects on recidivism, as well as to estimate the effect of comprehensive criminal justice reforms that change both incarceration and conviction decisions.

We have three main findings. First, we find that the effects of criminal justice contact differ substantially for defendants facing misdemeanor versus more serious felony charges. Incarceration for a felony offense reduces future crime through incapacitation, albeit at a relatively high monetary cost. In contrast, misdemeanor defendants' future outcomes are much more profoundly affected by conviction than by incarceration. Convictions result in *higher* levels of future crime and—consistent with a key role for criminal records—these effects are larger for defendants without serious prior convictions.

Second, in this setting 2SLS does not always reliably estimate the treatment effects of interest: a researcher instrumenting for conviction with judge assignment would dramatically overstate the increases in crime that result from conviction. We do, however, find that the exclusion and monotonicity violations are small when instrumenting for incarceration.

Finally, we investigate a series of possible criminal justice reforms and show there is scope for policies to reduce punitiveness while improving public safety. Increasing judges' leniency in determining guilt for misdemeanor defendants would reduce future crime and involvement with the criminal justice system. However, these gains disproportionately accrue to defendants accused of more serious misdemeanor offenses, and so policies that focus on this group would be even more beneficial.

We begin by examining what 2SLS allows us to learn about defendant outcomes under conviction and incarceration, relative to dismissal. The standard approach in the literature is to instrument for a single treatment of interest with judge assignment. We decompose the 2SLS estimand into its constituent effects and show that it will typically not be interpretable as a treatment effect because of the presence of exclusion violations: when judges vary in multiple margins of treatment, it is not possible to attribute the variation in outcomes across judges to a single margin. One possible solution is to instrument for both treatments simultaneously; however, as is well-known from prior work, the resulting estimand will typically be interpretable only if the effect of each of the treatments is the same for all complier groups.

These findings motivate us to develop a more flexible monotonicity assumption and corresponding estimation approach to credibly recover heterogeneous effects in settings with more than two treatments. Past work can be divided into two broad categories. The first has used either information on individuals' fallback options (Kirkeboen, Leuven and Mogstad, 2016), or instruments with multiple dimensions (Mountjoy, 2022) to credibly restrict the possible compliance types. We lack the data for the first approach and appropriate instruments for the second, since we observe only the judge assignment.

A second literature focused on the case of discrete instruments has instead imposed enough restrictions on how individuals' treatment choice responds to their instrument assignment that the share of each compliance type is identified from the data (Imbens and Angrist, 1994; Heckman and Pinto, 2018).[1] Our institutional setting, however, necessitates particular forms of flexibility in the admissible response patterns. The law requires judges to consider different, unrelated factors when deciding whether to convict a defendant at trial and whether to incarcerate a guilty defendant during the sentencing process. It is likely that judges evaluate each of these different dimensions of choice in a heterogeneous way, which would lead to pairs of judges with one set of compliers that move from conviction to incarceration, and another set that moves from incarceration to dismissal.

To account for this needed flexibility, we develop a *latent monotonicity* assumption that accommodates the two-way flows required in our setting. We show this latent monotonicity nests the compliance groups available in both ordered (Angrist and Imbens, 1995) and unordered monotonicity (Heckman and Pinto, 2018), two assumptions that are commonly used

[1] With ordered models, identifying the shares of each compliance group is less pressing. Angrist and Imbens (1995) show that while the shares of individuals moving between each pair of treatments is not identified, 2SLS still returns a proper weighted average of the effect of each treatment relative to the next-most-severe, which they call an average causal response.

in settings with discrete instruments. Latent monotonicity also overcomes a novel problem we identify that can arise with many judges. As the number of judges with varying treatment propensities grows, the ordered and unordered monotonicity assumptions converge to a much more restrictive, single-index model of treatment choice (Heckman and Vytlacil, 2005; Rivera, 2023); a model based on latent monotonicity does not.

The cost of our flexible monotonicity assumption is that we are required to develop a novel estimation approach. We translate our assumptions on potential outcomes into an equivalent threshold model of judge decisions, and demonstrate that the parameters of this selection model—and thus, the size and even existence of each compliance group—are not identified from the data, making it unclear how one could estimate treatment effects. Our insight is that for each admissible selection model that is consistent with the observable judge treatment propensities, one can estimate marginal treatment response functions via simple linear regression and aggregate them up to recover a wide variety of 2SLS-weighted treatment effects and policy-relevant treatment effects (Brinch, Mogstad and Wiswall, 2017; Heckman and Vytlacil, 1999, 2005). Under a semiparametric assumption on the primitives that drive selection, we can tractably search across all of these admissible selection patterns. Since the true underlying selection pattern is not identified from the data, we take the union of these estimates to comprise the identified set for the treatment effect of interest (Kamat, Norris and Pecenco, 2023).

We estimate the model using data from the three biggest counties in Ohio, which encompass the cities of Cleveland, Columbus, and Cincinnati (Norris, Pecenco and Weaver, 2021). These felony and misdemeanor courts are broadly representative of the American criminal justice system during our study period, reaching from the early 1990s to 2016. Key for our purposes, defendants are randomly assigned to judges, who are responsible for conducting trials and approving any plea deals. We use our method to study the effect of conviction and incarceration—relative to each other and to dismissal—on the number of new charges and convictions in the 5 years following the focal case filing date.

The 2SLS approach suggests that incarceration reduces the number of future offenses by 0.362 over the following five years, a 22% reduction relative to the mean. It also suggests that felony conviction leads to increases in future charges of 0.519 (a 33% increase), although the overall effect of conviction for both misdemeanors and felonies is much closer to zero.

To explore the validity of the 2SLS estimates, we first use our model to estimate treatment effects stripped of bias resulting from exclusion and monotonicity violations. For greatest comparability with standard approaches we target the 2SLS-weighted treatment effects for *net compliers*, those individuals who are on average shifted into incarceration by the judge instruments. Averaging across felony and misdemeanor courts, we find that the 2SLS estimate using incarceration as the treatment accurately reflects the effect of incarceration relative to conviction—nearly all of the weight is on compliers who are moved from conviction to incarceration by the instruments, and the exclusion and monotonicity terms are small and statistically insignificant. In contrast, the 2SLS estimate for conviction (whether or not the defendant was also incarcerated) is plagued by important violations of exclusion, and does not provide a reliable guide to the effect of conviction on future crime.

Examining the estimates in more detail, we find that the treatment effects differ dramatically between misdemeanor and felony courts. For felony defendants, we find no evidence that conviction affects crime, but we do find that incarceration dramatically lowers future charges and future convictions, by $[-0.246, -0.202]$ and $[-.278, -0.233]$ respectively.[2] Consistent with other recent work (Norris, Pecenco and Weaver, 2021; Rose and Shem-Tov, 2021), the effect of incarceration on crime in each year tracks the effects on the number of days spent incarcerated in that year. We conclude that these results are mostly driven by incapacitation effects.

In contrast, we see no effect of misdemeanor incarceration, potentially because sentences are short and hence there is little scope for incapacitation. Conviction, however, increases recidivism by a considerable amount, particularly for defendants without a prior felony conviction. For this group, a conviction causes an additional 0.165 to 0.817 crimes to be committed over the subsequent five years. The effects are even larger on the number of offenses a defendant is *convicted* of over the next five years, consistent with future police officers, prosecutors and judges being less lenient towards individuals with longer criminal records.

Motivated by the 2SLS-weighted effects, we next use our method to directly estimate the effects of possible policies. We consider reforms that would marginally increase leniency in either the conviction dimension (such as a higher evidentiary standard, or a policy of not prosecuting marginal cases) or the sentencing dimension (such as a change to structured sentencing guidelines to make them more lenient, or more consideration of mitigating factors). Leniency in sentencing would increase future crime in felony courts through decreased incapacitation.[3] Leniency in misdemeanor convictions, however, would decrease future charges and convictions at essentially no cost. These benefits are particularly large for defendants charged with more serious offenses and who would also be incarcerated if convicted.[4]

Our analysis relates to several literatures. We first contribute to the extensive body of work using examiner assignment designs. Numerous papers have acknowledged that examiners often choose between more than two options and that this is a threat to causal identification, including in the specific setting of crime we focus on (Bhuller et al., 2020; Norris, Pecenco and Weaver, 2021; Mueller-Smith, 2015). In these cases, they note that 2SLS estimates may be interpreted as causal effects only under strong homogeneity assumptions such as constant treatment effects across individuals.[5] Our analysis, in contrast, shows how to estimate treatment and policy effects while allowing for heterogeneous treatment effects and flexible choice behavior by judges.

Furthermore, our method provides a blueprint for structural estimation of heterogeneous treatment effects in examiner designs that is strictly more general than existing work. As

[2] Throughout the paper we use square brackets to denote bounds.

[3] Incapacitation is an expensive way to reduce crime. We calculate that each averted crime costs between \$55,000 and \$105,000 in prison costs alone, even before accounting for other social impacts.

[4] We also consider larger policy changes that would eliminate either conviction or incarceration, and find that there might be even larger benefits from policy reforms that target non-marginal defendants.

[5] Bhuller and Sigstad (2022) provides a high-level condition on the compliance groups that restricts two-way flows and under which 2SLS delivers a positively-weighted average of heterogeneous treatment effects. However, this condition is not usually satisfied under standard models of judge decision-making, including the ones we consider in this paper.

we discuss in Section 5.4, existing approaches allow for heterogeneity in treatment effects by relying on either restrictive models of treatment assignment (Rivera, 2023) or through assumptions about the existence of special additional regressors that shift judge behavior in particular ways (Humphries et al., 2023). With respect to Rivera (2023), the bounds that arise from our method will necessarily include the estimates from their model. And while not required for identification, our method can accommodate the additional regressors used for identification in Humphries et al. (2023). This means that if there was a setting where these regressors existed, they could be used along with the judge assignment as a source of additional variation. In contrast to Humphries et al. (2023), however, our method can still be applied when the researcher is uncertain about the credibility of these additional assumptions.[6]

In developing these results, our work contributes more generally to the instrumental variable analysis of treatment effects in settings with multiple treatments. Our setup considers a generalization of the binary-treatment monotonicity assumption of Imbens and Angrist (1994) that is weaker than the other multiple treatment generalizations applicable to our setup, such as ordered monotonicity (Angrist and Imbens, 1995) and unordered monotonicity (Heckman and Pinto, 2018). In particular, our model allows for what would be traditionally considered defiers—defendants who move out of incarceration when assigned to a more severe judge—as well as two-way flows into and out of conviction. In this sense, our analysis is related to the recent literature that highlights the usefulness of non-standard monotonicity assumptions (e.g., Lee and Salanié, 2018; Mogstad, Torgovitsky and Walters, 2021; Mountjoy, 2022; Pinto, 2019), and is applicable in a wide variety of settings where examiners decide between multiple treatments.

Our approach to identifying treatment effects using our threshold model also differs from the standard approach to do so in multiple treatment setups under discrete variation of the instrument. Previous work has focused on parameterizations of the selection model that result in point identification (e.g., Hull, 2020; Kline and Walters, 2016). However, an implication of our flexible monotonicity assumption is that point-identifying restrictions for our selection model do not naturally arise. We instead allow identification of treatment effects under a more flexible, partially identified selection model. Our approach is also distinct from alternatives used in the literature to identify related parameters, which exploit continuous variation in the instrument (e.g., Heckman, Urzua and Vytlacil, 2008; Lee and Salanié, 2018; Mountjoy, 2022) or additional data (Kirkeboen, Leuven and Mogstad, 2016).

Finally, we contribute to a rapidly-expanding literature studying the effects of incarceration and conviction. Similarly to us, one contemporaneous paper has focused on estimating the effects of these two treatments simultaneously using felony court data (Humphries et al., 2023). Their 2SLS results suggest, as do ours, that incarceration decreases future crime while conviction increases it. However, their structural results diverge sharply from ours; they find that only conviction affects future charges, while we find that only incarceration affects future charges. While these differences might arise from the differences in settings, they also suggest

[6] As we discuss in Section 5.4, there are some differences between the latent monotonicity model we use and the multinomial choice model that is implemented in Humphries et al. (2023). However, our approach of partially identifying the first stage can be straightforwardly implemented in their model.

the potentially important role of our differing methodological approaches.

A larger literature has focused on estimating the effect of conviction and incarceration in isolation. Recent work focusing on felony courts has mostly found that incarceration leads to reductions in future crime,[7] but that prosecution and conviction for more minor crimes tends to increase recidivism (Agan, Doleac and Harvey, 2022; Mueller-Smith and Schnepel, 2021). It has been unclear whether these disparate findings are caused by differences in the treatment (conviction versus incarceration), type of offense (misdemeanor or felony), research design, or geographic location. We reconcile this disparate literature by replicating the qualitative pattern of effects by treatment and offense severity found across these different studies using a single unified method in a single location.

2 Background

2.1 Setting

This study uses data from the courts in the three largest counties in Ohio: Franklin County (containing Columbus), Cuyahoga County (containing Cleveland), and Hamilton County (containing Cincinnati). The state is broadly representative of the criminal justice system in the United States in terms of both incarceration and recidivism rates.[8] Cases are divided into Municipal courts which handle misdemeanor criminal and traffic offenses,[9] and Common Pleas courts, which oversee more serious felony cases.

Importantly for the design of this study, Ohio law mandates that most criminal cases are randomly assigned to judges. The random assignment is carried out by a computer program and done separately by court. Defendants with ongoing cases or who are on probation are excluded from this randomization, although this amounts to a minority of cases. We drop all non-randomly assigned cases from our sample, leaving the analysis sample weighted towards first-time and non-chronic offenders. We conduct all analysis using the identity of the first-assigned judge to account for any issues arising from the approximately 5% of cases who are transferred between judges often due to likely workload and scheduling issues. To restrict comparisons between the set of judges available at any given time, throughout the paper we include court-year fixed effects.

Our discussion of the guilt and sentencing decisions in this paper follows from the institutional context used in criminal proceedings. The first part of the criminal proceedings focuses on whether a defendant is guilty—a defendant can either be found or plead guilty, or the case can undergo some version of dismissal. This decision is made without regard to the possible sentence. For example, in Ohio jurors are instructed that they are not allowed to consider

[7] For example, some find decreased crime (Rose and Shem-Tov, 2021; Kuziemko, 2012; Norris, Pecenco and Weaver, 2021; Huttunen, Kaila and Nix, 2021; Bhuller et al., 2020), mixed results (Green and Winik, 2010; Estelle and Phillips, 2018; Loeffler, 2013; Harding et al., 2017), and increased crime (Mueller-Smith, 2015).

[8] See Norris, Pecenco and Weaver (2021) for more details.

[9] We focus on misdemeanor cases of 1st to 4th degree to focus on cases commonly considered criminal. This excludes traffic cases and minor misdemeanors, which are very low-level offenses such as noise ordinance complaints.

the possible punishments when deciding guilt. The second part of criminal proceedings determine sentencing. Information that is inadmissible in the first part of the criminal proceedings such as aspects of the defendant's criminal or social history, mitigating circumstances, and victims' inputs are able to be obtained, often through a pre-sentencing investigation. Judges' sentencing decisions are also not supposed to reflect their prior on the defendant's guilt. Consequently, the date of guilt determination and sentencing determinations are often different, although both decisions for low-level misdemeanors often occur on the same day given the limited sanctions involved.[10]

2.2 Data

We collect and match administrative court data. Court records are available starting in the early 1990s (exact date depending on the county) and contain the full case history, including charges, arraignment date, sentencing date and decisions (punishment type and sentence length), and defendant characteristics (name, date of birth, sex, race, and home address). We collected all available misdemeanor and felony case records from the three counties, totaling 2.6 million cases and 862,505 unique defendants.

These data are used to measure sentencing outcomes and judge assignment, as well as whether defendants engage in criminal activity in the future. The sample construction and matching procedure follows Norris, Pecenco and Weaver (2021) closely. We match defendants to future charges by name and date of birth. For each match, we block on date of birth, then measure name similarity by Jaro-Winkler distance. If there is a perfect match on name, we keep only that match. Failing that, we keep matches with a Jaro-Winkler score higher than 0.9 for both first and last name. This is a high threshold but allows some room for spelling and transcription mistakes. Name and date of birth are unique for the vast majority of defendants in our sample.[11]

2.3 Summary statistics

Table 1 summarizes the characteristics of defendants for the randomly assigned cases in this sample. Column (2) shows the defendants are disproportionately male (77%), have an average age of 32, and while the broader population of the counties is mostly white, the majority of defendants are black (61%). Drug and property crimes are the most common offenses (29% of cases each) with fewer for violent (19%), family (14%), and sex (6%) offenses. About 30% of cases have at least one charge that isn't well-described by these categories. Defendants have committed on average 2.2 previous offenses, although the distribution is heavily right-skewed. The median defendant has no past charges, while a defendant at the 90$^{\text{th}}$ percentile has 6 past charges and at the 99$^{\text{th}}$ percentile has 19.

Table A1 presents summary statistics separately by whether the defendant was charged

[10] For example, in one of the Common Pleas courts (Cuyahoga county) with relevant data on judgment and sentencing dates, 70% of plea dates and sentencing dates differ. When they differ, the median duration is 33 days.

[11] See Norris, Pecenco and Weaver (2021) for details.

with a felony or a misdemeanor. Felony defendants, who are charged in the Common Pleas courts, are more likely to have a previous charge and to be charged with a property or drug crime, although not a violent crime. They are much more likely to be convicted and incarcerated (87% and 29%, respectively) than misdemeanor defendants, who are charged in Municipal courts (52% and 10%, respectively). The form of these sanctions also differ; the average defendant in the felony court receives a sentence of approximately 200 days, as compared to 4 days in the misdemeanor court. Conditional on incarceration, the median defendant in the felony court is sentenced for 270 days while the median incarcerated defendant in the misdemeanor court receives a 15 day sentence. Therefore, analyzing the data separately by offense type is important due to both defendant and treatment heterogeneity.

Finally, we show evidence consistent with the quasi-random assignment of judges, an assumption required for the following analysis. Columns (3) and (4) in Table 1 show that the judges' incarceration and conviction severity are uncorrelated with observable characteristics of defendants (p=0.79 and p=0.28, respectively) conditional on stratifying fixed effects, providing one indication that judge assignment is likely to be uncorrelated with defendant unobservables.

3 IV interpretation and results

In this section we present simple 2SLS estimates of the effect of incarceration and criminal conviction on outcomes. We then discuss the possible challenges to identification and interpretation that arise from the multiple margins of treatment in our setting.

3.1 IV estimates

A common research design to estimate the effect of a treatment, such as incarceration or conviction, on future criminal outcomes is to instrument for the treatment with the identity of the assigned judge. This research design is typically justified by the monotonicity and exclusion conditions in Imbens and Angrist (1994), which assumes that moving from a lower to a higher treatment propensity judge moves defendants from other possible treatment states into the treatment state of interest, but not in any other directions. In this section, we present 2SLS estimates based on this research design.

To describe our regression specifications, it is useful to introduce some notation. For each individual, we observe Z, D, X, and Y, where $Z \in \mathcal{Z} = \{z_1, \ldots, z_J\}$ denotes the judge to which the individual is assigned, $D \in \mathcal{D}$ the individual's treatment, $X \in \mathcal{X}$ the case covariates, and $Y \in \mathcal{Y}$ the outcome of interest capturing future criminal behavior. Given our focus on the conviction and incarceration margins, we aggregate the set of judicial decisions to be $D = \{n, c, p\}$, where n denotes no conviction, c denotes conviction without incarceration, and p denotes conviction with incarceration.[12] We analyze treatments T_i that equal $D_p = \mathbb{1}[D{=}p]$,

[12] We pick p for the mnemonic with "prison," although not all incarcerated defendants technically go to prisons, which are run by the state. Those who are incarcerated on short sentences go to county-run jails.

indicating an individual is incarcerated, or $D_{cp} = \mathbb{1}[D \in \{c, p\}]$, indicating any form of criminal conviction.

For each of these treatments, we use the following 2SLS specification:

$$Y_{ix} = \beta^{\text{2SLS}} T_i + \mu_x + \varepsilon_{ix} \tag{1}$$

$$T_i = \sum_{j=1}^{J} \alpha_j \mathbb{1}[Z = z_j] + \phi_x + e_{ix} \tag{2}$$

where μ_x and ϕ_x are court-year fixed effects as required by the design.[13]

Table 3 reports the estimated effects of incarceration (Panel A) and criminal conviction (Panel B) on cumulative number of charges over the following five year period. If the Imbens and Angrist (1994) assumptions are satisfied, column (1) in Panel A shows that incarceration reduces the number of future charges by 0.38. Relative to the mean of 1.6 charges over the next 5 years, this is a consequential 23% decrease. The other columns show additional heterogeneity to contextualize future results. Columns (2) and (3) show the results separately in the felony and misdemeanor courts. Incarceration significantly reduces future charges by 0.48 for defendants in the felony court, while there is a marginally statistically significant ($p < 0.10$) reduction of 0.15 charges for misdemeanor defendants. Although there has been much recent interest in the potential impacts of criminal justice sanctions for individuals who have not previously been convicted of a felony, columns (4) and (5) show similar impacts for this population compared to the full population in both courts.

Turning towards the estimated effects of conviction in Panel B, column (1) shows little average effect, although this result masks important heterogeneity across offense types. Column (2) suggests that conviction in felony courts is highly criminogenic, increasing future charges by 0.52, while column (3) finds a statistically insignificant and smaller decrease in future charges arising from a misdemeanor conviction.[14] Columns (4) and (5) report similar results for the sample without a previous felony conviction.

The validity of the arguments in Imbens and Angrist (1994) and its extensions is based on the assumption of a single treatment. In our context, this translates to the requirement that the assignment to a higher treatment propensity judge makes all defendants weakly more likely to be assigned that treatment (monotonicity) and doesn't move any defendants between the non-focal treatments. As judges may affect both conviction and incarceration, this raises concerns on the causal interpretation of the above conclusions. A common attempt (e.g., Mueller-Smith, 2015; Bhuller et al., 2020; Norris, Pecenco and Weaver, 2021) to account for this concern is to run an augmented 2SLS specification where the second stage includes both

[13] Judge instruments are sometimes constructed as leave-one-out averages to ameliorate potential finite sample bias issues. Because there are many observations per judge this turns out to be empirically unimportant in our setting.

[14] The overall null effect across both courts despite the large estimated felony effect indicates that much of the variation across judges in conviction propensities occurs in misdemeanor courts.

treatments, and where each treatment is instrumented for with the judge assignment as follows:

$$Y_{ix} = \beta^{\text{2SLS}^*}_{incar}\,\mathbb{1}D_p + \beta^{\text{2SLS}^*}_{convic}\mathbb{1}D_{cp} + \mu_x + \varepsilon_{ix} \tag{3}$$

$$D_p = \sum_{j=1}^{J} \alpha^1_j \mathbb{1}[Z = z_j] + \phi^1_x + e^1_{ix} \tag{4}$$

$$D_{cp} = \sum_{j=1}^{J} \alpha^2_j \mathbb{1}[Z = z_j] + \phi^2_x + e^2_{ix}. \tag{5}$$

Panel C in Table 3 reports estimates of the coefficients $\beta^{\text{2SLS}^*}_{incar}$ and $\beta^{\text{2SLS}^*}_{convic}$. While the results for the impacts of incarceration and overall effects of conviction are unchanged, column (2) shows a somewhat weaker effect of felony conviction when controlling for incarceration. While these broadly similar results may support the empirical conclusions of the single treatment regressions, previous work has noted the challenges in interpreting this multiple-treatment specification. In the next section, we develop a framework for interpreting these estimates further.

3.2 When is the 2SLS estimand interpretable?

These 2SLS estimates are interpretable as treatment effects only under certain assumptions. To provide a basis for our analysis, and to highlight how the presence of multiple treatments makes the required assumptions more onerous, in this section we provide a simple decomposition of the 2SLS estimand into its constituent parts: the target parameter, exclusion violations, and monotonicity violations. We focus on the single-treatment 2SLS estimand in (1) when one instruments for incarceration, but a similar analysis applies if one instead instruments for conviction. We focus on the single-treatment estimand in order to provide an exact, analytical decomposition of the parameter most commonly studied in the literature. We further discuss the multiple-treatment 2SLS estimand at the end of this section.

We suppose that the observed decision and future criminal behavior are generated by the usual potential outcomes structure. Using $D(z)$ to denote the potential decision had the individual been assigned to judge $z \in \mathcal{Z}$, the observed decision is assumed to be given by

$$D = \sum_{z \in \mathcal{Z}} 1\{Z = z\} D(z)\ . \tag{6}$$

Similarly, using $Y(d)$ to denote the potential criminal behavior had the individual's treatment been $d \in \mathcal{D}$, the observed behavior is assumed to be given by

$$Y = \sum_{d \in \mathcal{D}} 1\{D = d\} Y(d)\ . \tag{7}$$

Given that judges are randomly assigned to individuals, we take the assigned judge to be an exogenous instrument in our analysis. Formally, we assume the following relation between the assigned judge and the underlying potential outcomes:

Assumption E *(Exogeneity) Z is jointly independent of $(Y(n), Y(c), Y(p), D(z_1), \ldots, D(z_J))$ conditional on X.*

In our setting, the covariates are the set of court-year indicators that control for the set of cases among which a given set of judges is randomly assigned.

The building block of our decomposition of the 2SLS estimand is the compliance group, which is defined by the intersection of the treatments that an individual is induced into by each judge $\{\bigcap_{z \in \mathcal{Z}} D(z)\}$. For any two treatments s and t, individuals in compliance group ℓ are on average either induced from s into t by the instruments, or from t into s. As a result, we define Δ_{ts} as the $s \rightarrow t$ effect for those compliance groups who are on average induced between s and t by the instruments, with the weights for each compliance group determined by how much they contribute to the 2SLS estimand. This leads to the following decomposition:

Proposition 1 *Under Assumption E, the 2SLS estimand in (1) with incarceration as the treatment can be decomposed as*

$$\beta^{2SLS}_{incar} = \phi_{pc}\Delta_{pc} + \phi_{pn}\Delta_{pn} + \underbrace{\phi_{cp}\Delta_{cp} + \phi_{np}\Delta_{np}}_{mono.\ violations} + \underbrace{\phi_{cn}\Delta_{cn} + \phi_{nc}\Delta_{nc}}_{exclusion\ violations} \tag{8}$$

where $\phi_{ts} \geq 0$ for all $t, s \in \{n, c, p\}$ and

$$\phi_{pc} + \phi_{pn} - \phi_{cp} - \phi_{np} = 1$$

Proof: see Appendix A1.

This proposition reveals that $\beta^{\text{2SLS}}_{incar}$ can be viewed as containing three components: the weighted effect of being in treatment p relative to the non-carceral treatments n and c, the weighted effect of being in the non-carceral treatments relative to p, and the weighted effect of being moved between the two non-carceral treatments.[15] Interpretation of $\beta^{\text{2SLS}}_{incar}$ as a treatment effect of incarceration relative to alternatives therefore requires eliminating the latter two components of the decomposition.

We discuss two natural ways to do so. Both require binary-treatment exclusion, the assumption that the instruments do not affect outcomes except by moving individuals between p and either n or c. This is accomplished by

Assumption EX *(Binary exclusion) For all individuals for whom there exist instrument values z, z' such that $D(z)=n, D(z')=c$, then $Y(n) = Y(c)$.*

Under Assumption EX, $\phi_{cn}\Delta_{cn} = \phi_{nc}\Delta_{nc} = 0$ and the exclusion violations are eliminated from (8). This condition can be satisfied in two ways: either there are no individuals who are moved between n and c by changes in instrument assignment, or receiving the treatment n rather than c does not affect the outcomes of the compliers between n and c.

[15] Interpretation of $\beta^{\text{2SLS}}_{convic}$ is analogous, but with different terms labelled as exclusion and monotonicity violations.

There are two main approaches to eliminating the monotonicity violations. First, one could assume that the treatment effects are constant across individuals. We can then can rewrite (8) as

$$\beta_{incar}^{\mathrm{2SLS}} = (\phi_{pc} - \phi_{cp})\Delta_{pc} + (\phi_{pn} - \phi_{np})\Delta_{pn}$$

which is a proper weighted average of treatment effects under the verifiable condition that $(\phi_{pc} - \phi_{cp}) \geq 0$ and $(\phi_{pn} - \phi_{np}) \geq 0$. However, a large literature has rejected the assumption of constant treatments across settings (Heckman and Vytlacil, 2005; Kirkeboen, Leuven and Mogstad, 2016). We can also assess the validity of this assumption in our setting. To do so, we perform Sargan's J test to test the over-identifying restrictions of our IV model, and strongly reject constant effects ($p{=}0$).

The other option is to make assumptions on how treatment responds to judge assignment. The traditional binary-treatment monotonicity assumption requires that no individual is moved from incarceration into one of the other treatments when she is assigned to a higher-incarceration-propensity judge; this implies that $\phi_{np} = \phi_{cp} = 0$. Therefore, under binary-treatment monotonicity and exclusion, β^{2SLS} is a proper weighted average of $c{\to}p$ and $n{\to}p$ effects.

Viewed through the lens of (8), however, the single-instrument IV assumptions become more difficult to believe. Exclusion requires believing that conviction has no effect on future outcomes, or that judges do not shift individuals between n and c. Audit studies have suggested that criminal records have strong effect on job call-back rates (Pager, 2003; Agan and Starr, 2018), which suggests they may also affect criminal behavior. The latter restriction is also in contradiction with the different standards for conviction and incarceration in our setting. Since incarcerated defendants might face the fallback treatment of either conviction (if the evidence they committed the crime was strong but the sentencing guidelines did not require incarceration) or not guilty (if the evidence they committed the crime was middling but the sentencing guidelines made incarceration mandatory if convicted), this suggests that different judges with similar incarceration propensities might differ in their fallback options, violating monotonicity.

These pitfalls of single-treatment 2SLS, unfortunately, are not overcome by instrumenting for both treatments simultaneously. In Appendix A2, we provide a detailed analysis of the 2SLS specification in (3), and show the estimands can be decomposed into compliance group-weighted effects across different treatment comparisons as in (8). These effects are a combination of the parameters from the single-treatment case. Despite trying to account for multiple margins of judge decision making, this specification does not generally recover positively-weighted combinations of treatment effects even under highly restrictive substitution patterns. Furthermore, even when judge-pair comparisons could be used to isolate margin-specific causal effects, 2SLS fails to isolate these comparisons. An important exception to these results is in the case of constant treatment effects, in which case the estimands are interpretable as causal estimates. We conclude that in the presence of treatment effect heterogeneity, a credible analysis requires both a more flexible monotonicity assumption and an estimation strategy that moves beyond 2SLS.

4 A more flexible choice model

Our goal is to provide a choice model that accommodates two key channels in how judges behave. First, for any pair of judges, defendants might be marginal between assignment to any two of the treatments. Concretely, a defendant could be (1) marginal between n (not guilty) and c (convicted but not incarcerated) if the evidence was middling and the crime was minor enough that incarceration was not a possible punishment; (2) marginal between c and p (incarcerated) if the evidence was strong that he committed the offense and the judge had sentencing discretion over whether to incarcerate; and (3) marginal between n and p if the evidence was middling and the crime was very serious.

Second, the choice model should reflect the quasi-ordered nature of the treatment, and allow for two-way flows into and out of treatments. In particular, we want to allow for changing assignment from judge z to z' to move some defendants from n to p, and others from p to c. This would arise if the more severe judge had a lower standard for guilt but a higher standard for incarceration.

We first discuss two existing approaches, ordered (Angrist and Imbens, 1995) and unordered monotonicity (Heckman and Pinto, 2018), and discuss their shortcomings for our setting. We also discuss a single latent index model and novelly link it to other popular approaches. Finally, we introduce our own flexible monotonicity assumption and discuss how it can be used to estimate treatment and policy effects.

4.1 Ordered and unordered monotonicity

In order to formally state these assumptions, it is useful to first introduce some additional notation to denote the conviction and incarceration decisions as binary variables. For each $z \in \mathcal{Z}$, let $D_{cp}(z) \equiv 1\{D(z) \in \{c, p\}\}$ denote an indicator for whether the individual is convicted, regardless of being incarcerated or not, and $D_d(z) \equiv 1\{D(z)=d\}$ denote an indicator for being in treatment d. This lets us succinctly define *joint monotonicity*, which we use as a building block to discuss the other monotonicity assumptions. For clarity we drop any conditioning on covariates, but each of these definitions can be understood as conditional on x's.

Assumption JM *(Joint Monotonicity) For each $z, z' \in \mathcal{Z}$, we have*

$$D_{cp}(z) \geq D_{cp}(z') \quad \text{or} \quad D_{cp}(z) \leq D_{cp}(z') \ ,$$
$$D_p(z) \geq D_p(z') \quad \text{or} \quad D_p(z) \leq D_p(z') \ .$$

This assumption presumes that judges can be ordered with respect to their decision to convict (whether or not they incarcerate), and separately by their decision to incarcerate. It is weaker than both ordered and unordered monotonicity. Specifically, ordered monotonicity (Angrist and Imbens, 1995) assumes that

Assumption OM *(Ordered Monotonicity) For each $z, z' \in \mathcal{Z}$, we have*

$$D_{cp}(z) \geq D_{cp}(z') \text{ and } D_p(z) \geq D_p(z') \ , \quad \text{or}$$
$$D_{cp}(z) \leq D_{cp}(z') \text{ and } D_p(z) \leq D_p(z') \ .$$

while unordered monotonicity (Heckman and Pinto, 2018) imposes that

Assumption UM *(Unordered Monotonicity) For each $z, z' \in \mathcal{Z}$, we have that Assumption JM is satisfied and*

$$D_p(z) \geq D_p(z') \text{ or } D_p(z) \leq D_p(z') \ .$$

While these assumptions may be appropriate for some settings, in many examiner designs they impose unrealistically strong assumptions on treatment assignment. To better see the implications of each of these assumptions, Table 2 displays the response types $(D(z), D(z'))$ between a pair of judges $z, z' \in \mathcal{Z}$. We focus here on the case where $D_{cp}(z) \geq D_{cp}(z')$ as it is imposed by all the assumptions.

The first row of Table 2 reveals that OM rules out any defier types—if one judge is more severe in terms of conviction, then each defendant must be more likely to be incarcerated by her. This is particularly problematic in our setting because judges are instructed to treat the conviction decision $D_{cp}(z)$ as separate from the sentencing decision (incarceration conditional on being convicted, $D_p(z)$), and so it seems likely that judges might differ in their severity across those two margins.

An alternative to ordered monotonicity is UM. However, the second set of rows in Table 2 reveals that UM also rules out some natural response patterns. These restrictions arise because UM disallows two-way flows into and out of a treatment to ensure identification of the complier shares. Two-way flows with respect to incarceration—i.e., pairs of judges with both $n \to p$ types and $p \to c$ types—are particularly natural given that judges might have different orderings of standards for conviction versus incarceration. UM also precludes two-way flows with respect to conviction: a pair of judges cannot have both $n \to c$ and $c \to p$ compliers. This substitution pattern would be expected if the judges do in fact behave like the treatments can be ordered. By disallowing two-way flows, therefore, UM disallows at least some of the substitution patterns that are likely to occur in our setting.

4.2 Single-index choice model

One parsimonious alternative to OM and UM is to require that treatment is determined by a single unobservable index (Heckman and Vytlacil, 2005; Rivera, 2023). A key advantage of this model is that it allows the researcher to estimate marginal treatment effects (MTEs) along this single dimension. However, we show in this section that the single index model rules out certain compliance patterns that are important in our setting. Furthermore, we provide a new result that demonstrates that the single-index model is closely related to both OM and UM; if some of the judges satisfy a particular condition relating to the shares of defendants assigned to each treatment, then all three models are identical. We take this as further evidence that OM and UM might be inappropriate for use in courts.

We begin by defining treatment assignment in the single-index model:

Assumption SI *(Single Index) For each judge $z \in \mathcal{Z}$, treatment is determined by*

$$D(z) = \begin{cases} n & \textit{if} \;\; g_1(z) < U \;, \\ c & \textit{if} \;\; g_2(z) < U \leq g_1(z) \;, \\ p & \textit{if} \;\; U \leq g_2(z) \;, \end{cases}$$

where $U \sim U[0,1]$ and $g_2(z) \leq g_1(z)$.

In the single-index model, defendants may be marginal between n and c or between c and p. However, they can only be marginal between n and p if the judge does not assign any defendants to c. In practice, since we don't observe any judges who don't assign anyone to c, this amounts to ruling out situations where judges are marginal between finding a defendant not guilty and incarcerating them. This, in turn, is at odds with accommodating defendants who face severe charges (and so would be incarcerated if convicted) but are marginal on whether they will be convicted.

We display the full set of allowable response types under SI in Table 2. The table reveals that SI also rules out two-way flows in and out of incarceration, another key substitution pattern that we seek to accommodate. We conclude that the single-index model is unlikely to be appropriate in our setting.

While SI may appear to be more restrictive than both OM and UM, there are in fact deep underlying similarities between the models. To demonstrate this, we focus on JM, which is weaker than both OM and UM. Building on Vytlacil (2002, 2006), we provide an index characterization of JM, then demonstrate an important condition under which it is equivalent to SI.

Proposition 2 *Assumption JM is equivalent to*

$$D(z) = \begin{cases} n & \textit{if} \;\; U_1 > g_1(z) \;, \\ c & \textit{if} \;\; U_1 \leq g_1(z), U_2 > g_2(z) \;, \\ p & \textit{if} \;\; U_1 \leq g_1(z), U_2 \leq g_2(z) \;, \end{cases} \tag{9}$$

for each $z \in \mathcal{Z}$, where $U_1, U_2 \sim U[0,1]$ and $g_1(z) \geq g_2(z)$, and

$$P[U_1 > g_1(z), U_2 \leq g_2(z)] = 0 \tag{10}$$

$$P[U_1 \leq g_1(z), U_2 > g_2(z)] = g_2(z) - g_1(z) \tag{11}$$

Furthermore, if for every $t \in [0,1]$ there exists $z \in \mathcal{Z}$ such that $g_1(z) = g_2(z) = t$, then JM is equivalent to SI.
Proof: see Appendix A3.

Corollary 2.1 *If for every $t \in [0,1]$ there exists $z \in \mathcal{Z}$ such that $P[D{=}n|Z{=}z] = t$ and $P[D{=}c|Z{=}z] = 0$, then OM and UM are equivalent to SI.*

Proposition 2 reveals that Assumption JM introduces a sequential threshold crossing structure on judge decisions: they first assign each individual a rank of not being convicted (U_1) and not being incarcerated (U_2), and then convict and additionally incarcerate convicted individuals with ranks below their thresholds. However, through (10) and (11), Proposition 2 reveals that Assumption JM also introduces an additional restriction on how judges allow individuals to differ in their rankings across the two decision margins. To better see the content of this restriction, Figure A4(a) graphically illustrates the inadmissible area of rankings in the case of a single judge who does not convict any defendants. Here we can see that while rankings such as (u_1', u_2') and (u_1', u_2') are permitted, those such as (u_1'', u_2') or (u_1', u_2'') that increase the rank of one decision margin relative to the other are not. This highlights that the restriction can imply that whenever a judge assigns an individual a high rank in one margin, they necessarily must do so in the other as well.

As illustrated in Figure A4(b), the restriction becomes stronger in the presence of more judges as the area of inadmissible rankings increases. Proposition 2 sharpens this observation when there is sufficient continuous variation in judges thresholds. It shows that in this case, Assumption JM imposes a homogeneous rank for the incarceration and conviction margins. This is a strong restriction on judge behavior. For example, consider an individual plausibly guilty of certain petty misdemeanor crimes. For such an individual, judges may assign a high rank of being convicted but not necessarily incarcerated. A homogeneous rank for the two decisions margin, however, rules out such realistic scenarios. Since JM is weaker than OM and UM, this also implies that these assumptions also converge to SI.

While there are no judges in our setting who assign no defendants to treatment c, the above proposition suggests that tests of the validity of SI may shed light on the reasonableness of models of ordered and unordered monotonicity. In Appendix A4, we show that under SI, for any characteristic X, the Wald estimands between two judges on the outcomes XD_d and treatments D_d for $d \in \{n, p\}$ are bounded. This follows because the characteristics X of treatment d individuals are exactly controlled by changes in the treatment share for these outcome moments in this model. We adopt a semiparametric test developed in Frandsen, Lefgren and Leslie (2023) designed for the case of judge comparisons and find that we reject this test for some covariates, indicating that the data appears to be inconsistent with SI.

4.3 Latent monotonicity

The inability of the previous monotonicity assumptions to simultaneously allow $n{\to}p$ and $p{\to}c$ types is at odds with our empirical setting. This issue arises from the logical connection between the conviction and incarceration variables on which they are based: if $D_{cp}(z) = 0$ then it must be that $D_p(z) = 0$. Solving this problem requires decoupling the incarceration and conviction decisions.

One natural way to do so is to apply the legal principle that sentencing happens through a different process than the determination of guilt, and hinges on different factors. Guilt is determined solely by the strength of the evidence that the defendant committed the crime; sentencing by the severity of the crime and the defendant's criminal record. Indeed, these decisions are often made by different people and at different times. In Ohio, as in most other states, in jury trials the jury determines guilt while the judge determines sentencing. And even in the case of a plea deal, sentencing usually occurs some time after the defendant has pleaded guilty and at a different hearing.

To implement this principle, we assume that judges would behave monotonically with respect to incarceration if they had convicted all defendants. More precisely, let $D^*_p(z)$ denote an indicator for whether an individual is potentially incarcerated by judge z in the hypothetical case where they were already convicted. This means that the potential treatment for whether the individual is incarcerated is given by $D_p(z) = D_{cp}(z)D^*_p(z)$. Importantly, observe here that $D^*_p(z)$ is defined independently of $D_{cp}(z)$ and so we can have $D^*_p(z) = 1$ even if $D_{cp}(z) = 0$. This allows the following monotonicity condition:

Assumption LM *(Latent Monotonicity) For each $z, z' \in \mathcal{Z}$, we have*

$$D_{cp}(z) \geq D_{cp}(z') \quad \text{or} \quad D_{cp}(z) \leq D_{cp}(z') \ ,$$
$$D^*_p(z) \geq D^*_p(z') \quad \text{or} \quad D^*_p(z) \leq D^*_p(z') \ .$$

Heuristically, this condition imposes that a judge's sentencing decision (whether or not to incarcerate) would not hinge on whether she thinks the appropriate guilt standard has been met, in line with the legal separation between these standards. LM allows for substantially more flexibility than the available alternatives. As can be seen in Table 2, in contrast to the other assumptions we consider LM allows for two-way flows in both conviction and incarceration. By making the incarceration monotonicity condition conditional on conviction—and breaking the mechanical dependence between the conviction and incarceration condition—we allow for a wider variety of compliance types. Furthermore, unlike OM and UM, LM continues to allow for rich compliance patterns no matter the distribution of judge treatment propensities. As in Proposition 2, we formally show this in the following proposition.

Proposition 3 *Given Assumption E, Assumption LM is equivalent to*

$$D(z) = \begin{cases} n & \text{if } \ U_1 > g_1(z) \ , \\ c & \text{if } \ U_1 \leq g_1(z) \ , \ U_2 > g_2(z) \ , \\ p & \text{if } \ U_1 \leq g_1(z) \ , \ U_2 \leq g_2(z) \ , \end{cases} \tag{12}$$

where $U_1 \sim U[0,1]$ and $U_2 \sim U[0,1]$, and $(g_1(z), g_2(z)) \in [0,1]^2$ are judge-specific thresholds. Let F denote the cumulative distribution function of the joint distribution of (U_1, U_2).
Proof: see Appendix A3.

The above proposition shows that Assumption LM is equivalent to imposing a two-stage threshold crossing equation for each judge's decision for an individual. Judges first decide whether to convict an individual or not, and then, conditional on conviction, they decide whether to incarcerate them or not. Panel A of Figure 1 provides a graphical description of the threshold structure in the space of the individual latent variables. U_1 can be interpreted as an individual's resistance to conviction, since judges first convict those with lower values, while U_2 can be interpreted as their resistance to incarceration since judges would first incarcerate those with lower values. Symmetrically, $g_1(z)$ and $g_2(z)$ can be respectively interpreted as a judge's level of severity along the conviction and incarceration margins, as those with higher values respectively convict and incarcerate more defendants.

While we favor this monotonicity assumption and the associated threshold model primarily because it shares a number of key features with the institutional context, we can also use our data to assess the model implications. As in the case of SI, Appendix A4 derives testable implications of this assumption on judge-pair Wald estimands over appropriately defined outcome moments and performs these tests. In contrast to SI, the data do not reject these implications, providing some support for the validity of LM.

5 Identification of treatment effects

The main identification challenge we face is that the parameters of the selection equation—and thus, the size and even existence of compliance groups—are not point-identified from the data. In this section we explain how we can nonetheless partially identify the first stage and then use polynomial marginal treatment response functions (Brinch, Mogstad and Wiswall, 2017) to partially identify a wide range of parameters of interest, including 2SLS-weighted treatment effects and policy-relevant treatment effects. However, the intuition behind our identification approach applies to a larger and more flexible class of models, including ones with different selection models and with nonparametric outcome equations (Kamat, Norris and Pecenco, 2023).

5.1 Indeterminacy of the first stage

The selection equations in (12) relate the parameters of the selection model to treatment decisions. They imply that for each judge, the treatment propensities are functions of those parameters:

$$P(D{=}n|Z{=}z, X{=}x) = 1 - g_1(z, x) \ , \tag{13}$$

$$P(D{=}c|Z{=}z, X{=}x) = g_1(z) - F_{U_1,U_2}(g_1(z, x), g_2(z, x)) \ , \tag{14}$$

$$P(D{=}p|Z{=}z, X{=}x) = F_{U_1,U_2}(g_1(z, x), g_2(z, x)) \ . \tag{15}$$

where we now add x as an argument to g to emphasize that our first-stage identification is conditional on covariates.

From (13), it is clear that $g_1(z, x)$ is directly identified from the data. However, (14) indicates that $g_2(z, x)$ depends crucially on the unobserved joint distribution of U, F.[16] Dif-

[16] By adding up of the treatment shares, (15) adds no additional information on g.

ferent distributions of F correspond to different values of $g_2(z,x)$ and thus different compliance patterns.

We illustrate the indeterminacy of the first stage in Panels B and C of Figure 1. The figure shows the response types between two judges when U is distributed as a normal copula with unknown correlation ρ. We assume that judge z incarcerates 10% of defendants and judge z' incarcerates 20%. They also differ in conviction shares ($P[D{=}c|Z{=}z] = 0.1$ and $P[D{=}c|Z{=}z'] = 0.6$), resulting in different types of compliers for different values of ρ. In Panel B, where $\rho{=}0$, 5% of the population is a $p{\rightarrow}c$ complier and there are no $c{\rightarrow}p$ compliers. In Panel C, where $\rho{=}0.8$, there are no $p{\rightarrow}c$ compliers. Instead, 2.9% of the population is a $c{\rightarrow}p$ complier. Therefore, neither the size nor existence of various compliance groups are determined by the observable judge treatment propensities.

Our challenge is to continue to learn about treatment effects despite non-identification of the first stage. To make progress on this front, we assume that U is distributed in a known parametric family, and in our application assume that it is a normal copula with correlation ρ. This substantially simplifies our task, because it means that for each value of ρ we can directly calculate $g(z)$ using (13)-(15) and thus identify the compliance groups between each pair of judges. We further assume that $\rho \geq 0$, so that defendants who have unobservable characteristics that make them more likely to be convicted also have unobservable characteristics that make them more likely to be incarcerated.

5.2 Parameters of interest

Our analysis exploits the fact that the parameters of interest can be expressed as known functions of the same primitives that determine the observed data. We take the primitive of the selection equation for the judge decisions to be the correlation ρ between the components of U, noting that ρ determines the judge thresholds g. Following Heckman and Vytlacil (1999, 2005), we take the primitives with respect to the outcomes to be the marginal treatment response (MTR) functions

$$m_d(u_1,u_2,x) \equiv E[Y(d)|U_1{=}u_1,U_2{=}u_2,X{=}x]$$

for $d \in \mathcal{D}$. We can then express our parameters of interest in terms of the MTRs and the unknown primitive of the selection equation, ρ. For example, a key object of interest is a version of the 2SLS estimand stripped of the exclusion and monotonicity violations present in (8). Specifically, we take the weighted average of the $n{\rightarrow}p$ and $c{\rightarrow}p$ effects:

$$\beta_{p,cn} = \omega_{pc}\Delta_{pc} + \omega_{pn}\Delta_{pn} \quad , \quad \omega_{st} = \frac{\phi_{st}}{\phi_{pc}+\phi_{pn}}$$

$$\Delta_{st} = \sum_{x\in\mathcal{X}} w_x \int [m_s(u_1,u_2,x) - m_t(u_1,u_2,x)w_{st}(u_1,u_2,x)dF(u_1,u_2)$$

where w_x is the share of cases with covariate x. The weights on the different compliance groups, w_{st} and ϕ_{st}, are defined in Appendix A1, where we also show that these weights are functions of ρ.

We also consider how to estimate the effect of particular policy reforms. For example, consider a policy that changes evidentiary burdens such that a judge becomes δ more lenient on the conviction margin. This will move some defendants from p to n, and others from c to n. Using $\mathcal{P}_{st}^{\delta}(z,x)$ to denote the types who are moved from treatment t to s by the policy change (e.g., $\mathcal{P}_{np}^{\delta}(z,x) = [g_1(z,x) - \delta, g_1(z,x)] \times [0, g_2(z,x)]$), the effect of the policy on outcomes is

$$\underbrace{\int_{u \in \mathcal{P}_{np}^{\delta}(z,x)} [m_n(u_1,u_2,x) - m_p(u_1,u_2,x)] dF(u_1,u_2)}_{p \to n \text{ effect for counterfactually incarcerated defendants}} + \underbrace{\int_{u \in \mathcal{P}_{nc}^{\delta}(z,x)} [m_n(u_1,u_2,x) - m_c(u_1,u_2,x)] dF(u_1,u_2)}_{c \to n \text{ effect for counterfactually convicted defendants}}$$

which is a function of both how many defendants are moved into n from p versus c, as well as the magnitude of the treatment effect for each of these groups. In Section 6.6 we report the effects of various policies that adjust judges' thresholds.

5.3 Regression based solution to identification problem

Just as the parameters of interest can be expressed as functions of the underlying primitives, so can the observed data. Our challenge is to learn which MTRs are consistent with the data, and hence, which values of the parameters of interest are also consistent with the data.

To do so, we assume the MTRs are polynomials in u_1 and u_2, where we allow the MTRs to differ across covariate cells:

$$m_d(u_1,u_2,x) = \sum_{k_1=0}^{K_{d1}} \sum_{k_2=0}^{K_{d2}} \alpha_{dk_1k_2x} u_1^{k_1} u_2^{k_2}$$

This allows us to write expected outcomes given judge assignment and treatment as a simple linear function of calculable regressors. In particular, we have that

$$\begin{aligned} E[Y \mid D{=}d, Z{=}z, X{=}x] &= \frac{1}{P_{dxz}} \int_{u \in \mathcal{U}_{dxz}(\rho)} m_d(u_1,u_2,x) dF(u_1,u_2) \\ &= \sum_{k_1=0}^{K_{d1}} \sum_{k_2=0}^{K_{d2}} \alpha_{dk_1k_2x} \underbrace{\int_{u \in \mathcal{U}_{dxz}(\rho)} \frac{u_1^{k_1} u_2^{k_2}}{P_{dxz}} dF(u_1,u_2)}_{\equiv h_{dk_1k_2xz}(\rho)} \end{aligned}$$

where P_{dz} is the likelihood that judge z assigns treatment d and $\mathcal{U}_{dxz}(\rho)$ is the rectangular area in the space of unobservables such that a defendant with covariate x and index u would receive treatment d if assigned to judge z.[17]

For each value of ρ, there is a different distribution of U (through F) as well as a different mapping between u and treatment (through $\mathcal{U}_{dz}(\rho)$). This implies a different relationship

[17] Specifically, $\mathcal{U}_{nxz}(\rho) = [g_1(z,x), 1] \times [0,1]$, $\mathcal{U}_{cxz}(\rho) = [0, g_1(z,x)] \times [g_2(z,x), 1]$, and $\mathcal{U}_{pxz}(\rho) = [0, g_1(z,x)] \times [0, g_2(z,x)]$.

between the selection indices u and outcomes, which is summarized by $h(\rho)$. It also suggests that for each value of ρ, this relationship can be recovered by a simple linear regression of outcomes on the calculable covariates h:

$$y_{idxz} = \sum_{k_1=0}^{K_{d1}} \sum_{k_2=0}^{K_{d2}} \alpha_{dk_1k_2x}(\rho) h_{dk_1k_2xz}(\rho) + \varepsilon_{idxz} \tag{16}$$

To implement our model, we allow the MTRs to vary across each of the six courts in our data, with the treatment-specific intercepts additionally varying at the year level. We view this as a middle ground between MTRs that do not vary with X—and thus use cross-court variation in h to identify the selection parameters—and allowing the MTRs to vary flexibly by court-year, which asks a lot of the data.[18]

This means that there are $6|D|((K_{d1}+1)(K_{d2}+1)-1)+|D||X|$ parameters, and $|D||X||Z|$ moments. There are therefore many more moments than parameters, and $\alpha(\rho)$ is point-identified for each value of ρ if the standard regression rank condition is satisfied.

As discussed in the previous section, our objects of interest are functions of the MTRs and ρ. Letting $\theta(\alpha, \rho)$ represent a generic parameter of interest, for any value of ρ we can calculate this parameter as $\theta\big(\alpha(\rho), \rho\big)$. However, since ρ is not identified from the data, there are a range of possible values of θ that will be consistent with the data. To summarize these values, we take the union of values of θ across different plausible values of ρ as our identified set (Kamat, Norris and Pecenco, 2023). Formally, the identified set is

$$\Theta = \Big\{\theta_0 : \theta_0 = \theta\big(\alpha(\rho), \rho\big) \text{ for some } \rho \in [0, 1]\Big\}.$$

Because h is a nonlinear function of ρ that must be calculated using numerical quadrature, in practice we estimate the identified set by calculating h only for values of ρ in the grid $[0, .0.2, ..., 1]$. For each ρ we then estimate (16) and use these estimates to calculate θ. We then take the smallest and largest values of θ across this grid as the lower and upper bounds on the parameter of interest.

5.4 Connection to other approaches

A number of recent papers have considered identification of multiple treatment effects in settings with either additional data on fallback options (Kirkeboen, Leuven and Mogstad, 2016) or instruments that vary in multiple dimensions (Mountjoy, 2022). We lack data on outside options and have only a single-dimensional discrete instrument—the assigned judge—and so cannot apply either of these approaches.

The Arteaga (2021) study of the effect of parental incarceration on child outcomes also considers the threshold selection model in (12). We make several additional contributions. First, we clarify which assumptions on potential outcomes give rise to this model, and connect

[18] The assumption of some separability on the MTRs is common practice in the applied MTE literature as this expands the region of identification of the outcome functions (Cornelissen et al., 2016). However, we emphasize that we do not use separability to identify the first stage.

it to the OM and UM choice models. Second, because of data limitations Arteaga (2021) uses the model only to motivate 2SLS regressions of outcomes on instrumented incarceration among the sample of convicted defendants, controlling for judges' conviction propensity. In contrast, we show how the same choice model can be used to estimate a variety of additional treatment effects, including 2SLS-weighted effects of conviction relative to both incarceration and case dismissal, as well as other policy-relevant treatment effects.

Our approach shares some similarities with Humphries et al. (2023), which also considers a setting with three treatments and examiner instruments. Their idea is that one might be able to use tools from the industrial organization literature to recover a multi-valued instrument from information on treatment propensities, and then apply Mountjoy (2022) to point-identify the treatment effects. Since Mountjoy (2022) depends on continuous variation—and judges are discrete—it is not clear how Mountjoy (2022) can be applied. More fundamentally, as discussed in Appendix A5, we show that their approach relies on the existence of additional regressors that vary within judge and affect decisions in a very particular separable manner. It is not clear what regressors might satisfy this unusual condition. It also means that in the usual best-case scenario of unconditionally randomly assigned judges and no additional covariates—or of judge effects that are nonseparable in the covariates—their model is not identified.

One other difference between our methods that could potentially lead to differences in results is the choice model. Rather than using (12), they adopt a multinomial choice model. This difference, however, is not a fundamental distinction between our respective approaches. Our strategy of partially identifying a semiparametric first stage and using the associated moments to estimate MTRs could be applied equally well with a multinomial choice model. If there was reason to believe that a covariate satisfied the separability condition they require, it could be added to the first stage and used as an additional source of variation. Even in the absence of these special covariates, however, our model would continue to be identified. As we discuss in Appendix A5, Humphries et al. (2023) would not be.

Finally, our approach is also closely related to the single-index model (Rivera, 2023), which is a special case of our model when $\rho = 1$. Our bounds therefore encompass the single-index estimates.

5.5 Implementation details

As previously discussed, we calculate g at the judge-year level. This means that for each value of ρ, there is a single set of thresholds g that satisfies the treatment shares in (13)-(15). We assume that the MTRs are partially separable, with a court-year-specific intercept and terms in u that vary by court. Since judges work in only one court, this means we exploit only within-court variation in judges' decision to identify the selection terms.

We assume that the c and p MTRs are second-order in each dimension of u, although we impose that the coefficient $\alpha_{d22x} = 0$ for numerical stability. Similarly, the n MTR is assumed to be second-order in u_1 but constant in the u_2 dimension, since all cross-judge variation in u_2 relevant to the n MTR is driven by changes in the u_1 dimension. We calculate the regressors $h(\rho)$ using numerical integration, then estimate $\alpha(\rho)$ and the corresponding object of interest for each $\rho \in [0, 0.2, ..., 1]$.We report the upper and lower bounds across ρ. Inference is conducted via the bootstrap separately for each value of ρ; we take the largest upper bound and smallest lower bound as (conservative) edges to the confidence interval for the parameter of interest.

We study two main types of parameters. First, we study the 2SLS-weighted net complier effects discussed in Section 5.2. To maintain comparability with the 2SLS results in Section 3, we use a result from Blandhol et al. (2022) to calculate the weights on the compliance groups implied by (1)-(2), and report effects using these weights. This means, for example, that if the binary 2SLS assumptions for instrumenting for incarceration were satisfied, our estimated parameter $\beta_{p,cn}$ would be equal to the 2SLS estimate. Second, we study a number of policy-relevant treatment effects, which imply different weights on the effects of each compliance group.

6 Results

6.1 Treatment effects vary over the range of admissible selection models

As discussed in the previous section, the observed information on judge treatment propensities is not sufficient to point-identify the selection equation. As a result, many objects of interest are only partially identified. We illustrate this in Figure 2, which shows the 2SLS-weighted effect of incarceration relative to conviction (Δ_{pc} from (8)) on the number of charges filed over the next five years for different values of ρ. When ρ is small, incarceration reduces the number of future charges by about 0.23. As ρ gets larger, however, and the selection model gets closer and closer to SI, the estimated effects decline, approaching -0.25. For ρ=1, which corresponds to the single-index model, however, the estimated effect shoots up to -0.20. Since the data do not provide any guidance on which ρ is the correct one, we conclude that Δ_{pc} is between -0.25 and -0.20, consistent with incarceration decreasing the number of future crimes through incapacitation.

6.2 Model fit

We approximate the MTRs using the flexible polynomials discussed in Section 5.5. These functions impose that the potential outcomes are smooth in the indices that govern selection, and place implicit restrictions on the slopes of potential outcomes across the space of u.

Before we proceed, we assess whether these MTRs accurately approximate the 2SLS estimates $\widehat{\beta}^{\text{2SLS}}_{incar}$. For each ρ, we denote the estimated MTRs $\widehat{\alpha}$ and the corresponding model-based incarceration 2SLS estimate as $\beta^{\text{2SLS}}_{incar}(\widehat{\alpha})$, and bootstrap the null distribution of $\widehat{\beta}^{\text{2SLS}}_{incar} - \beta^{\text{2SLS}}_{incar}(\widehat{\alpha})$ using i.i.d. draws from the set of cases.

Figure A1 shows the estimated p-values for models of both the number of charges and an indicator for any charge after 5 years over the range of ρ. The 2SLS estimates tend to be larger in absolute value than their structural analogs; for example, the 2SLS estimate for any charge is -0.058 while the structural estimate $\beta^{\text{2SLS}}_{incar}(\widehat{\alpha})$ with ρ=0 is -0.044.

Since the choice model is not correct for all values of ρ, we expect that the model might reject for some values. We find that the p-values are close to zero for ρ=1, consistent with the SI model being too restrictive. For other values of ρ, the p-values are higher, in the range of 0.20 for the binary outcome and 0.03 for the continuous outcome. We conclude that the polynomials do a reasonably good job of approximating the underlying MTRs.

6.3 Decomposing the 2SLS estimands

Except in restrictive choice models, the existence of multiple treatments poses complications for the interpretation of binary 2SLS estimates. As we discuss in Section 3.2, the incarceration 2SLS estimate, for example, is a weighted combination of the effect of incarceration relative to both conviction and dismissal, as well as monotonicity and exclusion violations. While researchers commonly make assumptions that preclude some or all of these effects, there is typically no empirical guide to evaluating their existence and magnitude.

Our model allows us to directly estimate each of these components and evaluate the threat that exclusion and monotonicity violations pose to interpretation. The following equation shows our decomposition of the incarceration 2SLS estimate on the number of charges over the five years following case filing using (8). For each treatment effect and weight, we report the upper and lower bounds on the model estimates across ρ, except for the model-based 2SLS estimand, for which we report the ρ=0 estimate.

$$\underbrace{-0.241}_{\beta^{\text{2SLS}}_{incar}} = \underbrace{[0.977,\ 1.000]}_{\phi_{pc}}\underbrace{[-0.246,\ -0.202]}_{\Delta_{pc}} + \underbrace{[0.000,\ 0.054]}_{\phi_{pn}}\underbrace{[0.220,\ 2.523]}_{\Delta_{pn}} + \underbrace{[-0.039,\ 0.013]}_{\text{mono. violations}} + \underbrace{[-0.003,\ 0.005]}_{\text{exclusion violations}} \tag{17}$$

The equation shows that 2SLS does a remarkably accurate job of estimating a causal effect of incarceration. It also reveals which effect we estimate: for each of the different admissible selection models, the weight is almost entirely on Δ_{pc}, the effect of incarceration relative to conviction. The weight on the $n \rightarrow p$ effect is smaller than 0.054 across all values of ρ, and so the weighted effect of incarceration relative to the alternatives $\beta_{p,cn}$, at $[-0.213, -0.202]$, almost perfectly coincides with the estimated $c{\rightarrow}p$ effect of $[-0.246, -0.202]$. Furthermore, the exclusion and monotonicity terms are tightly bounded around zero, meaning that the 2SLS estimate almost exactly reflects $\beta_{p,cn}$.

We also decompose the conviction 2SLS estimand into its constituent parts. This estimate results from instrumenting for conviction (either p or c) with judge indicators. In our choice model there are no $c \rightarrow n$ or $p \rightarrow n$ defiers, and so the conviction 2SLS estimand can be decomposed into a weighted combination of (1) the treatment effect of p relative to n, (2) the treatment effect of c relative to n, and (3) exclusion violations between c and p:

$$\underbrace{0.189}_{\beta^{\text{2SLS}}_{convic}} = \underbrace{[0.000,\ 0.237]}_{\phi_{pn}}\underbrace{[0.486,\ 2.897]}_{\Delta_{pn}} + \underbrace{[0.768,\ 1.005]}_{\phi_{cn}}\underbrace{[-0.189,\ 0.017]}_{\Delta_{cn}} + \underbrace{[0.111,\ 0.311]}_{\text{exclusion violations}} \tag{18}$$

This decomposition reveals that the conviction 2SLS estimate of 0.189 does not accurately reflect the causal effect of conviction on future criminal behavior, nor any other causal effect. While the majority of the weight is on $n \to c$ compliers, in contrast to the 2SLS estimate these effects are mostly negative, ranging from -0.189 to 0.017. The effect for the $n \to p$ group is larger, at $[0.486, 2.897]$, but the weight on these compliers is small enough that the combined $\beta_{pc,n}$ effect is no larger than 0.078.[19] The exclusion violations are bounded between 0.111 and 0.311 and statistically significant. Thus, a naive use of 2SLS would overstate the increases in crime resulting from conviction.

The decompositions also reveal the importance of the fallback option in assessing the effect of incarceration. While incarceration relative to conviction reduces the amount of future crime, incarceration relative to case dismissal increases it. In the next sections we explore this pattern in more detail.

6.4 Effects of conviction and incarceration on future criminal behavior

Table 4 presents the effects of incarceration and conviction on future outcomes for criminal defendants, combined over compliance groups using the weights implicit in the 2SLS regression of outcomes on incarceration instrumented with judge indicators. We use these weights as they correspond to the complier-weighted population most commonly focused upon in the literature. We generalize our findings to additional populations by directly studying policy-relevant treatment effects in the following section.

The first column of Panel A, which is identical to (17), reports the effect of conviction and incarceration on the number of times the defendant is charged with a new offense over the five years following case filing. Column (1) reveals that nearly all of the variation in the judge instruments shifts defendants between c and p; the weight on the $c \to p$ effect is at least 0.977 across values of ρ. This allows a precise estimate of between 0.202 and 0.246 future charged offenses averted for each marginal incarceration.

In contrast, the treatments that lead to a conviction, Δ_{pn} and Δ_{cn}, are relatively imprecisely estimated, and we can't reject that there is no effect of either treatment. While this might suggest that convictions are not important determinants of criminal justice outcomes, this is driven by two different factors: relatively low precision for felony cases caused by limited across-judge variation in conviction propensity, and smaller effects of a conviction for defendants who already have a criminal record. When we focus on the populations more likely to be affected by a conviction, and for whom we have more statistical power, we will see more precisely estimated and deleterious effects of conviction.

Column (2) of Table 4 reports the effects of conviction and incarceration for felony defendants. At least 99.1% of the weight is on the $c \to p$ effect, and for this group the table reveals that incarceration reduces the number of future charges by between 0.345 and 0.379 over the five years following filing. These effects accrue during the first several years following the case, when the sentence is still ongoing: Figure A3 plots the effects of incarceration on both days spent incapacitated in each year following case filing as well as the effect on the number of

[19] To see this, note we can rewrite $\beta^{2SLS}_{convic} = \beta_{pc,n} +$ exclusion violations.

charges filed against the defendant up until each year. By the third year following filing, incarceration results in only 25 days spent in prison, and has averted about 0.4 crimes. Over the next four years, there is no substantive effect on the number of days spent incapacitated, and no further change in the effect on number of crimes. We conclude that incapacitation rather than changes to post-release behavior likely explain nearly all of the effects of incarceration relative to conviction, consistent with the large incapacitation effects documented in prior work (Norris, Pecenco and Weaver, 2021; Rose and Shem-Tov, 2021).[20] We also study the effect of incarceration on future convictions (Panel B), and find substantively similar results.

Unfortunately, the effects of conviction on future behavior for felony defendants are relatively imprecisely estimated because of limited variation in conviction propensities, and we cannot reject a null of no effect on future charges (Panel A) or convictions (Panel B). One interesting exception though, is the effect on sentence length. $n \to p$ compliers spend on average between 530 and 4,486 days in prison as a result of incarceration. In contrast, $c \to p$ compliers spend between 383 and 392 additional days, consistent with them being convicted of more minor offenses and being marginal between probation and incarceration. The $n \to p$ compliers, in contrast, appear to be marginal between being found not guilty and being convicted of a serious crime with a long sentence.

In column (3) of Table 4 we turn to misdemeanor defendants. Misdemeanors are substantially less serious than felony offenses, and the sentence for marginally-incarcerated $c \to p$ and $n \to p$ defendants is bounded to only $[29, 30]$ and $[11, 29]$ days respectively. Figure 3 shows the effect on days incapacitated for each of these groups, and confirms that the effects are modest. As a result, there is substantially less scope for incapacitation to affect defendants' future behavior. Indeed, the bounds for the effect of incarceration rather than conviction on the number of future charges contains zero, and we can rule out increases or decreases of larger than 0.20 five years after case filing.

Despite the limited incapacitation effects for misdemeanors, convictions could still affect individuals' future outcomes. A criminal record might make it harder to gain employment (Pager, 2003), and since police officers and prosecutors will be able to see the record of convictions, future criminal justice system involvement might be more likely to result in charges. We explore this by examining the $n \to p$ and $n \to c$ effects, and find evidence that convictions actually increase future crime. For $n \to c$ compliers, conviction has no detectable effect on future charges (the bounds include zero), but increases the number of future convictions by $[0.222, 0.559]$.[21] For $n \to p$ compliers, incarceration relative to dismissal increases future charges by $[0.564, 4.106]$, and future convictions by $[0.787, 4.988]$. These effects arise almost exclusively through the conviction channel; we estimate that the $n \to c$ effect for the $n \to p$ group is $[1.88, 3.93]$ with a confidence interval of $(0.84, 6.40)$.[22] We interpret this to mean that conviction on more serious misdemeanor offenses is a key driver of future criminality.

[20] Given that the Δ_{pc} effect on sentence length is a precisely estimated $[383, 392]$ days, the Δ_{pc} estimate in Table 4 implies that each year of incarceration for the marginal felony defendant averts between 0.321 and 0.361 new offenses.

[21] The lower bound is statistically significant only at the 10% level.

[22] For each compliance group ℓ, the $n \to p$ effect Δ^{ℓ}_{pn} can be decomposed into $\Delta^{\ell}_{pc} + \Delta^{\ell}_{cn}$. We then aggregate these up with the compliance group specific $n \to p$ weights.

The positive effects of conviction on future crime have several interesting implications. First, the high rates of misdemeanor conviction in the US imply that the increases in crime and future conviction we find may be particularly important. Second, the larger effects on future convictions than on future charges are consistent with police, prosecutors, and judges treating defendants more harshly in *future* cases as a result of a past conviction, and suggest that initial inequities in criminal justice contact can lead to persistently differential treatment. Finally, our results suggest that recent work finding increased criminal offending as a result of misdemeanor prosecution may be driven by criminal conviction rather than other aspects of the charging process, which our results hold fixed.

More evidence on the central role of convictions in misdemeanor cases comes in column (5) of Table 4. In this column we restrict to misdemeanor defendants who have never been convicted of a felony offense, and so might be more profoundly affected by a conviction.[23] Consistent with this, we see that the lower bounds on the $n \to p$ and $n \to c$ effects are larger for this group than for all misdemeanor defendants. Weighting by the relative size of the two groups, conviction causes an additional $[0.165, 0.817]$ charges to be filed over the next five years. This is composed of $[0.851, 3.336]$ for defendants who are induced from not guilty into incarceration by the judge instruments, and by $[0.165, 0.351]$ for compliers who are induced into a non-carceral sentence. The effects on the number of future convictions are slightly larger for both compliance groups, again consistent with more punitive behavior from future criminal justice officials.

Table A2 reports incarceration and conviction effects using the implied weights from a regression instrumenting for conviction rather than for incarceration. This aggregation allows us to study a population more responsive to differences across judges with varying conviction propensities and provides a sense of heterogeneity in results when compared with the results using incarceration weights. Comparing effects on charges within 5 years, of most note, the effect of felony incarceration relative to conviction is now a positive though insignificant [0.069, 0.133], and we can reject this estimate overlaps with the incarceration 2SLS-weighted estimate. This finding highlights important heterogeneity in the effects of incarceration, and is inconsistent with constant treatment effects in this population. While we also see differential responses in effects of incarceration relative to no conviction, the effects of conviction relative to not guilty appear to be consistent across the complier populations.

6.5 Discussion

One stark pattern in the recent economics of crime literature has been the differing effects of incarceration and conviction on future criminal behavior. As we discuss in the introduction, recent studies have typically found that incarceration (relative to conviction) on felony charges reduces future crime through incapacitation, while other studies have found that prosecution (relative to dismissal) and conviction (relative to a deferred adjudication that leaves the defen-

[23] We also examine the effect of criminal justice sanctions for never-previously-convicted *felony* defendants in column (4). However, given the imprecision of our estimated conviction effects for felony defendants, we cannot rule out even relatively large changes in behavior.

dant without a criminal record) on more minor charges increases future crime. Thus, although both of these treatments (dismissal versus conviction, and conviction versus incarceration) increase the intensity of criminal justice contact and the associated degree of social stigma, they have opposite effects on recidivism. Whether this difference in effects is broadly generalizable across the United States, or stems from methodological or external validity distinctions, is an important unexplained puzzle in the economics of crime.

Our study provides a simple explanation for this pattern of results. Using a single research design and setting, we replicate the qualitative patterns of the prior literature. Conviction on minor offenses—like those studied in Agan, Doleac and Harvey (2022) and Mueller-Smith and Schnepel (2021)—increases future crime, possibly through the effect of a criminal record. Incarceration on felony charges decreases future crime through incapacitation. In other words, the differences across studies may be driven by the differential effects of conviction and incarceration for minor versus major offenses, rather than differences in research design or setting.

6.6 Policy effects

In the previous section, we reported effects of conviction and incarceration for individuals weighted by their response to the changes in the judge instruments. While these weights provide a useful benchmark, they do not capture populations likely to be represented by any policy changes (Heckman and Vytlacil, 2005). In this section, we directly study the effects of policies that change judge behavior.[24] Crucially, by estimating the effects of changing conviction and incarceration behavior in one setting, we can identify the impacts of policies that realistically induce responses along both of these margins.

We focus on two types of policy reforms. First, *local* changes make small reductions to judge thresholds $g_1(z)$ and $g_2(z)$. We view these as approximating what would happen if judges became more lenient with respect to conviction or sentencing, respectively. Greater conviction leniency could arise from a higher evidentiary standard, or greater willingness of prosecutors to drop cases where they viewed the evidence as marginal. Sentencing leniency could arise from reforms to sentencing guidelines that made probation the presumptive sentence for a wider range of defendants. We estimate the effects of greater conviction leniency by adjusting each judges' g_1 threshold by 0.01, reassigning treatment with the new threshold, and using the estimated MTRs to predict outcomes under the counterfactual policy. Similarly, we simulate the effect of greater incarceration leniency by decreasing g_2 by 0.01 for each judge, then estimating outcomes under the new treatment assignment.

Second, we study *global* policy changes that eliminate either conviction or incarceration. We estimate the effect of these policies by reducing the g_1 (respectively, g_2) threshold to zero for each judge, then calculating outcomes under the new treatment assignment with the estimated MTRs. Importantly, these policy effects do not take into account general equilibrium responses, and so may overstate the benefits of these policies. Nonetheless, they help illustrate the possible effects of larger, non-marginal policy changes.

[24] One could imagine using this framework to define similar policies affecting other criminal justice system actors such as prosecutors in a model of their behavior.

Panel A of Table 5 reports the effects each of the policy changes for felony defendants. Consistent with the 2SLS effects, both marginal increases in sentencing leniency as well as banning incarceration end up increasing the number of offenses committed by the defendants over the following fives years. The first row reports that the marginal defendant who would be spared incarceration by increasing sentencing leniency will commit an additional 0.356 to 0.66 crimes as a result of the policy change, and be convicted in an additional 0.256 to 0.609 cases. Similarly, banning incarceration would reduce the incarceration rate from 28.9% to 0% and increase the number of future charges by $[0.085, 0.295]$ per defendant (or $[0.294, 1.021]$ per affected person) even before accounting for any general equilibrium effects.

Although these results mean that expanding sentencing leniency would increase crime, they do not consist of a full cost-benefit analysis. In particular, prisons are expensive; a year of incarceration in Ohio costs approximately $26,500 (Mai and Subramanian, 2017). We estimate that the marginal incarceration in this policy change is for $[500, 513]$ days, implying a cost per averted crime of somewhere between $55,002 and $104,621.[25]

We also study the effect of expanded conviction leniency for felony defendants. However, consistent with the wide standard errors and limited effects of conviction for this population, we can make no firm conclusions about the likely effect of expanded conviction leniency for this population.

Panels B and C of Table 5 study the effect of increased leniency for misdemeanor defendants and never-previously-convicted misdemeanor defendants, respectively. In line with the small incapacitation effects we observed in Section 6.4, the bounds on the effect of increased sentencing leniency staddle zero for both populations. While not directly-crime reducing, this type of leniency would still reduce jail populations without large increases in future crime.

Expanded leniency in conviction seems somewhat more promising. The second rows of Panel B and C report the effect of marginally increasing conviction leniency for misdemeanor defendants. They reveal that on average it would have no effect on the number of future charges for the overall misdemeanor defendant (bounds of -0.112 to 0.215), and slightly (although not statistically significantly) reduce charges by 0.022 to 0.247 for the never-previously-convicted population. The benefits are slightly larger in terms of avoiding future convictions; the marginal never-previously-convicted beneficiary would see 0.224 to 0.489 fewer convictions as a result of this policy change.

These benefits are somewhat smaller than would be expected from the 2SLS estimates. A closer examination of the compliance patterns reveals why. For each population, 80-100% of the beneficiaries of increased conviction leniency would otherwise be convicted but not incarcerated. As we discussed in Section 6.4, the benefits of case dismissal are smaller for the $n \to c$ population than the $n \to p$ population. This substantially reduces the benefits of expanding leniency across the board, and suggests that more effective policy reforms would focus on defendants who would otherwise be incarcerated, since the benefits disproportionately accrue to this group.

We also study the effect of larger increases in conviction leniency. While these estimates

[25] $26500 \times ([500,\ 513]/365)/[0.356,\ 0.660] = [55002, 104621]$.

do not account for changes in general deterrence, they suggest that the effect of case dismissal for non-marginal defendants are substantially higher. For example, dismissing all cases would reduce the number of future convictions by $[0.858, 0.869]/0.53 = [1.62, 1.64]$ for the average affected misdemeanor defendant, compared to $[0.052, 0.449]$ for the defendant affected by a marginal increase in leniency. Policies that encourage diversion for misdemeanor defendants, and particularly those with a shorter criminal record, therefore have the potential to decrease crime and future involvement with the criminal justice system.

6.7 Robustness

Our main results report effects after 5 years. We choose this horizon because it is long enough that several years have elapsed since most defendants have been released (see Figure 3), but short enough that the sample size remains large. As a robustness exercise, we also estimate these effects after 7 years. Table A3 contains the incarceration 2SLS-weighted effects (analogous to Table 4), and Table A4 reports policy effects after 7 years (analogous to Table 5). While the edges of the bounds are sometimes slightly smaller or larger, the substantive conclusions are unchanged.

7 Conclusion

Estimating the causal effect of criminal justice sanctions is difficult due to non-random assignment of treatments. Examiner designs use variation from randomly assigned judges as instrumental variables to study the effect of a particular sanction, such as incarceration, on individual outcomes. However, simple 2SLS models cannot account for judges choosing between three or more treatments—such as dismissal, conviction and incarceration—thereby biasing estimates.

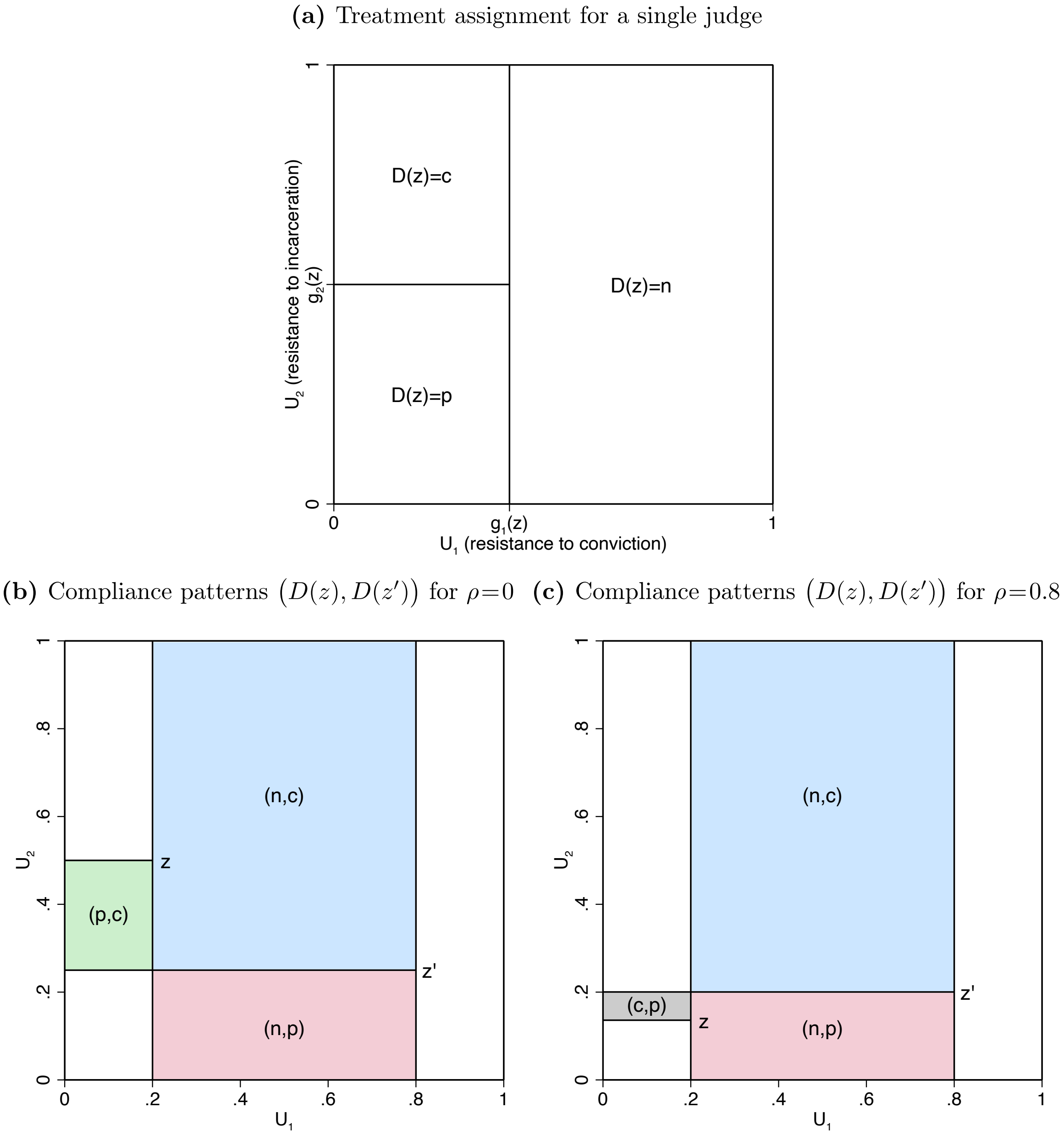
We build a new framework to handle these and other similar situations that feature discrete instruments and multiple treatments. We introduce a novel monotonicity assumption and an equivalent selection model, which allows us to recover 2SLS-weighted combinations of treatment effects stripped of monotonicity and exclusion violations, and to go beyond these estimands to estimate the component treatment effects and to extrapolate to well-defined alternative policies. Our approach, which requires only examiner instruments for identification, allows for more flexible substitution patterns—and thus, more credible estimates—than existing alternatives.

We use this model to study the effect of conviction and incarceration in the three largest counties in Ohio. We reconcile a string of recent and seemingly-contradictory results in the economics of crime literature studying these treatments. Consistent with prior work, we find that incarceration decreases future crime through an incapacitation effect, while misdemeanor convictions *increase* subsequent criminal justice involvement. Our contribution is to show that these results hold in a single setting and using a single research design, assuaging concerns about the internal and external validity of prior work. We also highlight the important differences in the effect of sanctions in misdemeanor and felony courts, emphasizing the importance of further work on this topic.

Most substantively, this paper has important implications for policy. We find that courts could implement reforms that both increase leniency and decrease crime, particularly if they target misdemeanor defendants with short criminal records and who face relatively serious charges.

Figures

Figure 1: Illustration of threshold crossing model for judge decisions

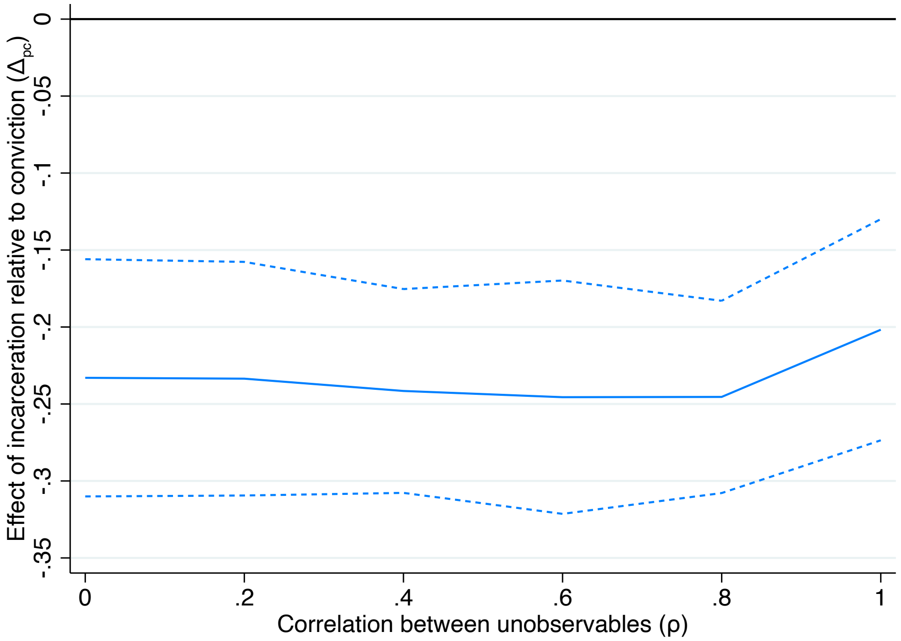


(a) Treatment assignment for a single judge

(b) Compliance patterns $\big(D(z), D(z')\big)$ for ρ=0 **(c)** Compliance patterns $\big(D(z), D(z')\big)$ for ρ=0.8

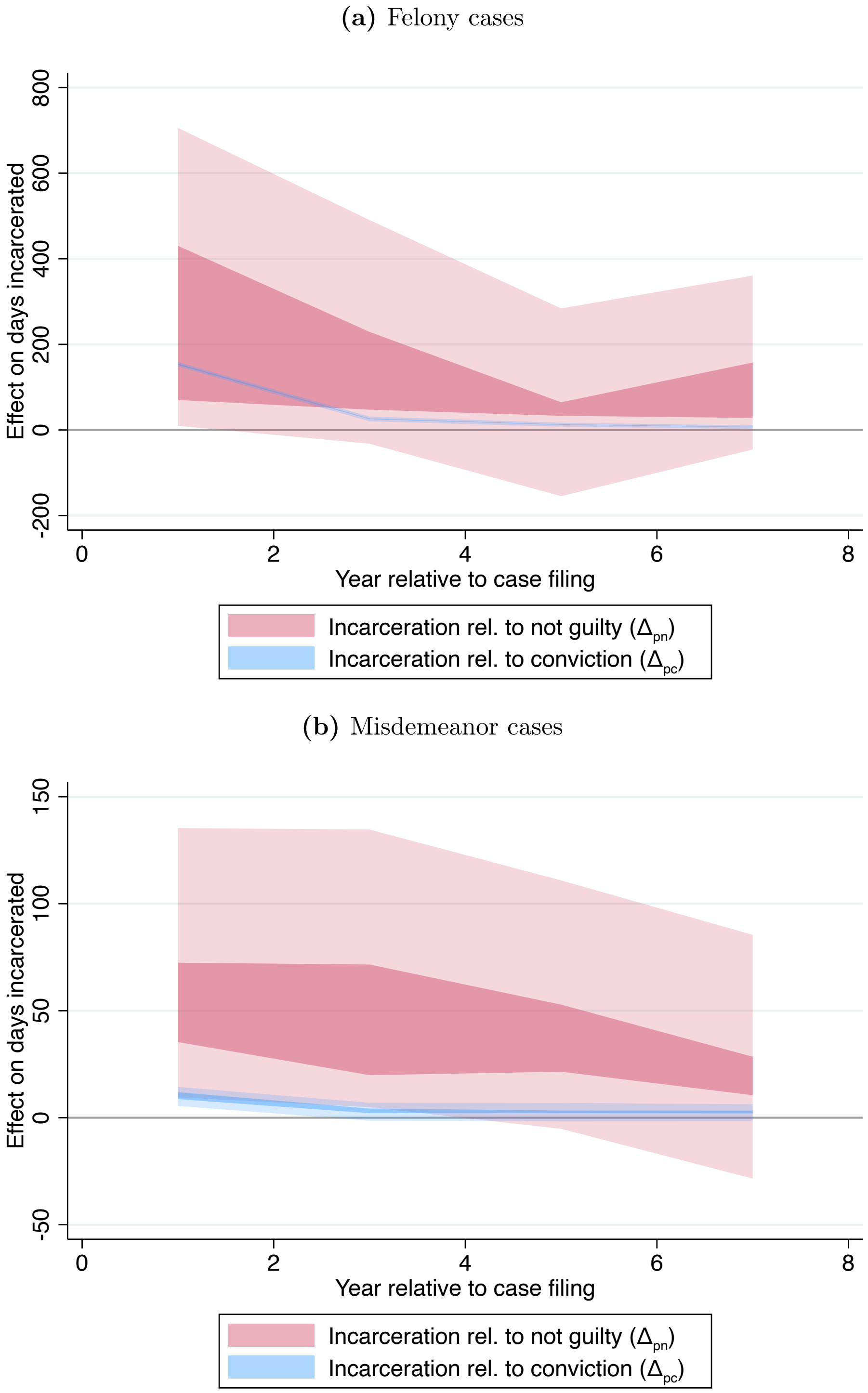
Panel A illustrates treatment assignment decisions for a single judge. Panels B and C show the compliance patterns $\big(D(z), D(z')\big)$ for judges z and z' for $\rho \in [0, 0.8]$. Judge z assigns 80, 10, and 10% of defendants to treatment n, c, and p, respectively. Judge z' assigns 20, 60, and 20%. As ρ changes so does the share of each response group; in particular there are (p, c) compliers for $\rho = 0$ but not $\rho = 0.8$. Similarly there are (c, p) compliers for $\rho = 0.8$ but not $\rho = 0$. The white regions in Panels B and C denote values of U where changing the judge from z to z' does not affect the realized treatment.

Figure 2: Effect of incarceration relative to conviction on 5-year number of charges, for different values of ρ



This figure shows the estimated effect of incarceration relative to conviction on the number of subsequent charges over the five years following the focal case filing. We display these estimates for $\rho \in [0, 0.2, 0.4, 0.6, 0.8, 1]$.

Figure 3: Effect of incarceration on days spent incarcerated, by year following case filing



(a) Felony cases

(b) Misdemeanor cases

This figure shows estimated effects of incarceration relative to not guilty (Δ_{pn}) and incarceration relative to conviction only (Δ_{pc}) on the number of days spent incarcerated 1, 3, 5, and 7 years after case filing. The darker areas denote the range of estimates arising from choice models with selection parameters $\rho \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$. The lighter areas denote the edge of 95% confidence intervals for the endpoints.

Tables

Table 1: Defendant characteristics and judge severity

	N	Mean	Incar. severity	Convic. severity
Male	633,591	.77	.0084	.037*
			(.008)	(.022)
White	638,411	.39	-.00077	-.023
			(.010)	(.026)
Age	638,684	31.99	-.14	-.71
			(.224)	(.566)
Drug crime	638,448	.29	-.016	.031
			(.010)	(.023)
Violent crime	638,448	.19	.0074	-.0036
			(.008)	(.021)
Property crime	638,448	.29	.013	-.036
			(.010)	(.023)
Sex crime	638,448	.05	.0033	.0011
			(.005)	(.012)
Family crime	638,448	.14	-.00097	.028
			(.006)	(.020)
Other crime	638,448	.28	-.0097	.013
			(.010)	(.024)
Sentence (days)	626,655	99.66	3	5.7
			(3.688)	(6.314)
Log sentence	623,742	3.00	.014	-.066
			(.025)	(.076)
Number of previous charges	638,684	2.17	-.088	.29
			(.079)	(.202)
Joint p-value			.79	.28

Columns (1-2) report the number of observations and sample means corresponding to this characteristic, respectively. Columns (3-4) report the coefficient from a regression of the characteristic on judge mean incarceration and conviction severity, respectively. Joint p-value comes from an F-test of joint significance of the characteristics on the instrument. Controls include year by court fixed effects. Cases may include multiple charges of different types, so the sum of types of charges sums to more than 1. Charge sentence measures offense severity by calculating the leave-out average sentence for the most serious charge. Standard errors clustered at the individual level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2: Response types $(D(z), D(z'))$ between a pair of judges $z, z' \in \mathcal{Z}$ with $D_{cp}(z) \geq D_{cp}(z')$

Assumption	Ordering	$(D(z), D(z'))$								
		(n,n)	(n,c)	(n,p)	(c,n)	(c,c)	(c,p)	(p,n)	(p,c)	(p,p)
Ordered	$D_p(z) \geq D_p(z')$	✓			✓	✓		✓	✓	✓
Unordered	$D_c(z) \geq D_c(z'),\ D_p(z) \geq D_p(z')$	✓			✓	✓		✓		✓
	$D_c(z) \geq D_c(z'),\ D_p(z) \leq D_p(z')$	✓			✓	✓	✓			✓
	$D_c(z) \leq D_c(z'),\ D_p(z) \geq D_p(z')$	✓				✓		✓	✓	✓
Single Index	$g_2(z) < g_1(z')$	✓			✓			✓	✓	✓
	$g_1(z') < g_2(z) < g_2(z')$	✓			✓	✓			✓	✓
	$g_2(z') < g_2(z)$	✓			✓	✓	✓			✓
Latent	$D_p^*(z) \geq D_p^*(z')$	✓			✓	✓		✓	✓	✓
	$D_p^*(z) \leq D_p^*(z')$	✓			✓	✓	✓	✓		✓

Treatments are n (not guilty or dismissed), c (convicted but not incarcerated), and p (incarcerated). Ordered refers to Assumption OM, and Unordered refers to Assumption UM, Single Index refers to Assumption SI and Latent refers to Assumption LM. Ordering refers to the additional monotonicity conditions imposed by the assumptions in addition to $D_{cp}(z) \geq D_{cp}(z')$, which is imposed by all four assumptions (in SI this implies that $g_1(z) \leq g_1(z')$). Note that the ordering $D_c(z) \leq D_c(z')$, $D_p(z) \geq D_p(z)$ doesn't logically exist when $D_{cp}(z) \geq D_{cp}(z)$.

Table 3: 2SLS effects on cumulative charges

				Never prev. convicted	
	All	Felony	Misdemeanor	Felony	Misdemeanor
	(1)	(2)	(3)	(4)	(5)
Panel A: Effect of incarceration					
Incarceration (D_p)	-0.362***	-0.484***	-0.131*	-0.391***	-0.033
	(0.043)	(0.052)	(0.075)	(0.076)	(0.089)
Dependent mean	1.616	1.581	1.653	0.977	0.933
Observations	638,684	323,046	315,638	143,657	167,258
Panel B: Effect of conviction					
Conviction (D_{cp})	0.084	0.519**	-0.076	0.352	0.040
	(0.108)	(0.203)	(0.128)	(0.215)	(0.127)
Dependent mean	1.616	1.581	1.653	0.977	0.933
Observations	638,684	323,046	315,638	143,657	167,258
Panel C: Effect of both					
Incarceration (D_p)	-0.363***	-0.475***	-0.128*	-0.382***	-0.031
	(0.043)	(0.053)	(0.075)	(0.078)	(0.090)
Conviction (D_{cp})	0.088	0.315	-0.054	0.151	0.036
	(0.109)	(0.208)	(0.129)	(0.221)	(0.127)
Dependent mean	1.616	1.581	1.653	0.977	0.933
Observations	638,684	323,046	315,638	143,657	167,258

This table reports IV estimates of the effect of incarceration, conviction, and both on cumulative charges up to 5 years post filing. Columns are split by sample, with column (1) including all cases, column (2) including felony cases, column (3) including misdemeanor cases, column (4) including felony cases for defendants with no prior felony convictions, and column (5) including misdemeanor cases for defendants with no prior felony convictions. The endogenous variables are instrumented with the judge identity and all specifications include court-year fixed effects. Standard errors in parentheses and clustered at individual level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4: Effects of conviction and incarceration on future criminal justice outcomes

				Never prev. convicted	
	All	Fel.	Misd.	Fel.	Misd.
	(1)	(2)	(3)	(4)	(5)
		Panel A: Number of charges over next 5 years			
Incarceration rel. to conviction (Δ_{pc})	[-0.246, -0.202]	[-0.379, -0.345]	[-0.021, 0.071]	[-0.333, -0.265]	[-0.037, 0.024]
	(-0.329, -0.139)	(-0.477, -0.261)	(-0.131, 0.185)	(-0.459, -0.174)	(-0.177, 0.151)
Incarceration rel. to not guilty (Δ_{pn})	[0.220, 2.523]	[-0.998, 0.018]	[0.564, 4.106]	[-4.237, -0.068]	[0.851, 3.336]
	(-0.465, 5.511)	(-6.442, 4.446)	(0.049, 6.601)	(-11.949, 3.475)	(0.248, 6.272)
Conviction rel. to not guilty (Δ_{cn})	[-0.185, 0.071]	[-0.303, -0.086]	[-0.124, 0.151]	[-0.187, 0.015]	[0.165, 0.351]
	(-0.325, 0.228)	(-0.720, 0.185)	(-0.314, 0.402)	(-0.475, 0.273)	(-0.043, 0.604)
		Panel B: Number of convictions over next 5 years			
Incarceration rel. to conviction (Δ_{pc})	[-0.278, -0.233]	[-0.364, -0.323]	[-0.149, -0.063]	[-0.272, -0.190]	[-0.073, -0.013]
	(-0.384, -0.144)	(-0.472, -0.221)	(-0.329, 0.115)	(-0.398, -0.098)	(-0.249, 0.159)
Incarceration rel. to not guilty (Δ_{pn})	[0.331, 3.054]	[-1.260, -0.063]	[0.787, 4.988]	[-0.954, 0.279]	[1.174, 3.898]
	(-0.606, 6.714)	(-7.815, 5.294)	(0.008, 8.575)	(-9.144, 7.237)	(0.153, 7.643)
Conviction rel. to not guilty (Δ_{cn})	[0.105, 0.358]	[-0.379, -0.032]	[0.222, 0.559]	[-0.139, 0.089]	[0.412, 0.629]
	(-0.093, 0.564)	(-0.854, 0.243)	(-0.024, 0.866)	(-0.468, 0.344)	(0.170, 0.902)
Weight on $c \rightarrow p$ effect	[0.977, 1.000]	[0.991, 1.000]	[0.947, 1.000]	[0.989, 1.000]	[0.966, 1.000]
Weight on $n \rightarrow p$ effect	[0.000, 0.054]	[0.000, 0.037]	[0.000, 0.086]	[0.000, 0.043]	[0.000, 0.074]
Weight on $n \rightarrow c$ effect	[0.065, 0.088]	[0.030, 0.045]	[0.131, 0.169]	[0.042, 0.058]	[0.124, 0.151]

This table reports treatment effects of conviction and incarceration, aggregated using the weights from a 2SLS regression with incarceration as the treatment and judge dummies as the instruments. MTRs are approximated by a second-degree polynomial in u_1 and u_2 as specified in Section 5.5.

Table 5: Policy effects after 5 years

	Change in treatment shares			Effects on outcomes	
	n	c	p	N. charges	N. conv.
			Panel A: Felony defendants		
Incarceration leniency ($g_2 \downarrow$)	0.000	[0.009, 0.010]	[-0.010, -0.009]	[0.356, 0.660]	[0.256, 0.609]
				(0.238, 0.953)	(0.122, 0.937)
Conviction leniency ($g_1 \downarrow$)	0.010	[-0.010, -0.007]	[-0.003, 0.000]	[-0.032, 0.195]	[-0.013, 0.128]
				(-0.307, 0.405)	(-0.319, 0.410)
No incarceration ($g_2 = 0$)	0.000	0.289	-0.289	[0.085, 0.295]	[0.102, 0.270]
				(-0.019, 0.355)	(-0.010, 0.342)
No conviction ($g_1 = 0$)	0.874	-0.585	-0.289	[-9.848, -9.799]	[-4.721, -4.662]
				(-20.402, 0.731)	(-19.115, 9.672)
			Panel B: Misdemeanor defendants		
Incarceration leniency ($g_2 \downarrow$)	0.000	[0.005, 0.010]	[-0.010, -0.005]	[-0.265, 0.190]	[-0.174, 0.383]
				(-0.400, 0.389)	(-0.346, 0.667)
Conviction leniency ($g_1 \downarrow$)	0.010	[-0.010, -0.008]	[-0.002, 0.000]	[-0.112, 0.215]	[-0.449, -0.052]
				(-0.305, 0.378)	(-0.676, 0.150)
No incarceration ($g_2 = 0$)	0.000	0.103	-0.104	[-0.038, 0.030]	[-0.033, 0.047]
				(-0.054, 0.055)	(-0.055, 0.086)
No conviction ($g_1 = 0$)	0.530	-0.427	-0.104	[-0.395, -0.386]	[-0.869, -0.858]
				(-0.962, 0.188)	(-1.484, -0.231)
			Panel C: Never-convicted misdemeanor defendants		
Incarceration leniency ($g_2 \downarrow$)	0.000	[0.005, 0.010]	[-0.010, -0.005]	[-0.193, 0.123]	[-0.186, 0.252]
				(-0.349, 0.351)	(-0.377, 0.561)
Conviction leniency ($g_1 \downarrow$)	0.010	[-0.010, -0.008]	[-0.002, 0.000]	[-0.247, -0.022]	[-0.489, -0.224]
				(-0.481, 0.146)	(-0.769, -0.018)
No incarceration ($g_2 = 0$)	0.000	0.087	-0.087	[-0.021, 0.010]	[-0.018, 0.025]
				(-0.036, 0.032)	(-0.037, 0.053)
No conviction ($g_1 = 0$)	0.528	-0.441	-0.087	[-0.447, -0.436]	[-0.731, -0.715]
				(-0.980, 0.088)	(-1.266, -0.163)

Table reports the effects of a number of policy changes on recidivism. We analyze marginal changes, which shift judges' thresholds g by 0.01, as well as global changes. The change in treatment shares is the change from the given policy. The change in outcomes is rescaled by the number of defendants whos' treatment is affected by the policy change for the marginal changes to assist in readability. Bounds are in square brackets and the outer edges of 95% confidence intervals in parentheses.

References

Agan, Amanda, Andrew Garin, Dmitri Koustas, Alex Mas, and Crystal Yang. 2023. "The Impact of Criminal Records on Employment, Earnings, and Tax Outcomes."

Agan, Amanda, and Sonja Starr. 2018. "Ban the box, criminal records, and racial discrimination: A field experiment." *The Quarterly Journal of Economics*, 133(1): 191–235.

Agan, Amanda Y, Jennifer L Doleac, and Anna Harvey. 2022. "Misdemeanor prosecution." National Bureau of Economic Research.

Angrist, Joshua D, and Guido W Imbens. 1995. "Two-stage least squares estimation of average causal effects in models with variable treatment intensity." *Journal of the American statistical Association*, 90(430): 431–442.

Arnold, David, Will Dobbie, and Crystal S Yang. 2018. "Racial bias in bail decisions." *The Quarterly Journal of Economics*, 133(4): 1885–1932.

Arteaga, Carolina. 2021. "Parental Incarceration and Children's Educational Attainment." *The Review of Economics and Statistics*, 1–45.

Berry, Steven T, and Philip A Haile. 2023. "Nonparametric Identification of Differentiated Products Demand Using Micro Data."

Bhuller, Manudeep, and Henrik Sigstad. 2022. "2SLS with multiple treatments." *arXiv preprint arXiv:2205.07836.*

Bhuller, Manudeep, Gordon B Dahl, Katrine V Løken, and Magne Mogstad. 2020. "Incarceration, recidivism, and employment." *Journal of Political Economy*, 128(4): 1269–1324.

Blandhol, Christine, John Bonney, Magne Mogstad, and Alexander Torgovitsky. 2022. "When is TSLS actually late?" National Bureau of Economic Research.

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

Cornelissen, Thomas, Christian Dustmann, Anna Raute, and Uta Schönberg. 2016. "From LATE to MTE: Alternative methods for the evaluation of policy interventions." *Labour Economics*, 41: 47–60.

Court Statistics Project. 2018. "State Court Caseload Digest." *Court Statistics Project.*

Dobbie, Will, Jacob Goldin, and Crystal S Yang. 2018. "The effects of pre-trial detention on conviction, future crime, and employment: Evidence from randomly assigned judges." *American Economic Review*, 108(2): 201–240.

Doyle, Joseph J, John A Graves, Jonathan Gruber, and Samuel A Kleiner. 2015. "Measuring returns to hospital care: Evidence from ambulance referral patterns." *Journal of Political Economy*, 123(1): 170–214.

Estelle, Sarah M, and David C Phillips. 2018. "Smart sentencing guidelines: The effect of marginal policy changes on recidivism." *Journal of public economics*, 164: 270–293.

Frandsen, Brigham, Lars Lefgren, and Emily Leslie. 2023. "Judging judge fixed effects." *American Economic Review*, 113(1): 253–277.

Green, Donald P, and Daniel Winik. 2010. "Using random judge assignments to estimate the effects of incarceration and probation on recidivism among drug offenders." *Criminology*, 48(2): 357–387.

Gross, Max, and E Jason Baron. 2022. “Temporary stays and persistent gains: The causal effects of foster care.” *American Economic Journal: Applied Economics*, 14(2): 170–199.

Harding, David J, Jeffrey D Morenoff, Anh P Nguyen, and Shawn D Bushway. 2017. “Short-and long-term effects of imprisonment on future felony convictions and prison admissions.” *Proceedings of the National Academy of Sciences*, 114(42): 11103–11108.

Heckman, James J, and Edward 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, James J, and Edward Vytlacil. 2005. “Structural equations, treatment effects, and econometric policy evaluation 1.” *Econometrica*, 73(3): 669–738.

Heckman, James J, and Rodrigo Pinto. 2018. “Unordered monotonicity.” *Econometrica*, 86(1): 1–35.

Heckman, James J., Sergio Urzua, and Edward Vytlacil. 2008. “Instrumental Variables in Models with Multiple Outcomes: the General Unordered Case.” *Annales d’Économie et de Statistique*, , (91/92): 151–174.

Hull, Peter. 2020. “Estimating hospital quality with quasi-experimental data.” *Available at SSRN 3118358*.

Humphries, John Eric, Aurelie Ouss, Kamelia Stavreva, Megan T Stevenson, and Winnie van Dijk. 2023. “Conviction, Incarceration, and Recidivism: Understanding the Revolving Door.”

Huttunen, Kristiina, Martti Kaila, and Emily Nix. 2021. “The Punishment Ladder: Estimating the Impact of Different Punishments on Defendant Outcomes.”

Imbens, Guido W, and Joshua D Angrist. 1994. “Identification and Estimation of Local Average Treatment Effects.” *Econometrica*, 62(2): 467–475.

Kamat, Vishal, Samuel Norris, and Matthew Pecenco. 2023. “Identification in Multiple Treatment Models under Discrete Variation.”

Kirkeboen, Lars J, Edwin Leuven, and Magne Mogstad. 2016. “Field of study, earnings, and self-selection.” *The Quarterly Journal of Economics*, 131(3): 1057–1111.

Kline, Patrick, and Christopher R Walters. 2016. “Evaluating public programs with close substitutes: The case of Head Start.” *The Quarterly Journal of Economics*, 131(4): 1795–1848.

Kling, Jeffrey R. 2006. “Incarceration length, employment, and earnings.” *American Economic Review*, 96(3): 863–876.

Kuziemko, Ilyana. 2012. “How should inmates be released from prison? An assessment of parole versus fixed-sentence regimes.” *The Quarterly Journal of Economics*, 128(1): 371–424.

Lee, Sokbae, and Bernard Salanié. 2018. “Identifying effects of multivalued treatments.” *Econometrica*, 86(6): 1939–1963.

Loeffler, Charles E. 2013. “Does imprisonment alter the life course? Evidence on crime and employment from a natural experiment.” *Criminology*, 51(1): 137–166.

Maestas, Nicole, Kathleen J Mullen, and Alexander Strand. 2013. “Does disability insurance receipt discourage work? Using examiner assignment to estimate causal effects of SSDI receipt.” *American economic review*, 103(5): 1797–1829.

Mai, Chris, and Ram Subramanian. 2017. “The price of prisons: Examining state spending trends, 2010-2015." *Vera Institute of Justice*, 7–8.

Mogstad, Magne, Alexander Torgovitsky, and Christopher R Walters. 2021. "The causal interpretation of two-stage least squares with multiple instrumental variables." *American Economic Review*, 111(11): 3663–98.

Mountjoy, Jack. 2022. "Community colleges and upward mobility." *Available at SSRN 3373801.*

Mueller-Smith, Michael. 2015. "The criminal and labor market impacts of incarceration." *Unpublished Working Paper*, 18.

Mueller-Smith, Michael, and Kevin Schnepel. 2021. "Diversion in the criminal justice system." *The Review of Economic Studies*, 88(2): 883–936.

Norris, Samuel, Matthew Pecenco, and Jeffrey Weaver. 2021. "The effects of parental and sibling incarceration: Evidence from ohio." *American Economic Review*, 111(9): 2926–63.

Pager, Devah. 2003. "The mark of a criminal record." *American journal of sociology*, 108(5): 937–975.

Pinto, Rodrigo. 2019. "Noncompliance as a rational choice: A framework that exploits compromises in social experiments to identify causal effects." *Unpublished working paper. University of California, Los Angeles.*

Rivera, Roman. 2023. "Release, Detain, or Surveil? The Effect of Electronic Monitoring on Defendant Outcomes." *Unpublished working paper.*

Rose, Evan K, and Yotam Shem-Tov. 2021. "How does incarceration affect reoffending? Estimating the dose-response function." *Journal of Political Economy*, 129(12): 3302–3356.

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

Vytlacil, EJ. 2006. "Ordered discrete choice selection models: Equivalence, nonequivalence, and representation results." *Review of Economics and Statistics*, 88(3): 578–581.

Appendix

A1 Interpreting the 2SLS estimand

In this section we provide a decomposition of the 2SLS estimand into constituent effects. To maintain clarity we abstract away from covariates. However, they can be easily accommodated using results from Blandhol et al. (2022), which we use when estimating this decomposition on our data.

The building blocks of our analysis are the compliance groups defined by the intersection of their treatment assignment under each judge $\{\bigcap_{z\in\mathcal{Z}} D(z)\}$. For compliance group ℓ define the measure of that group as π_ℓ, and the average treatment effect of s versus n as $\Delta^\ell_{st} = E[Y(s) - Y(t) \mid C_\ell]$ where C_ℓ denotes membership in group ℓ. Using $D_{\ell j}$ as shorthand for $D(z_j)$ for compliance group ℓ, we decompose the binary incarceration 2SLS estimand as

$$
\begin{aligned}
\beta^{2\text{SLS}_{incar}} &= \sum_{j=1}^{J} \lambda_j \frac{E[Y|z_j] - E[Y|z_{j-1}]}{E[D_p|z_j] - E[D_p|z_{j-1}]} &&\text{(A1)}\\
&= \sum_{j=1}^{J} \frac{\lambda_j}{E[D_p|z_j] - E[D_p|z_{j-1}]} \Big(\sum_\ell \Delta^\ell_{pc} \mathbb{1}[D_{\ell j}{=}p, D_{\ell,j-1}{=}c]\pi_\ell + \sum_\ell \Delta^\ell_{pn} \mathbb{1}[D_{\ell j}{=}p, D_{\ell,j-1}{=}n]\pi_\ell +\\
&\qquad \sum_\ell \Delta^\ell_{cp} \mathbb{1}[D_{\ell j}{=}c, D_{\ell,j-1}{=}p]\pi_\ell + \sum_\ell \Delta^\ell_{cn} \mathbb{1}[D_{\ell j}{=}c, D_{\ell,j-1}{=}n]\pi_\ell +\\
&\qquad \sum_\ell \Delta^\ell_{np} \mathbb{1}[D_{\ell j}{=}n, D_{\ell,j-1}{=}p]\pi_\ell + \sum_\ell \Delta^\ell_{nc} \mathbb{1}[D_{\ell j}{=}n, D_{\ell,j-1}{=}c]\pi_\ell \Big)\\
&= \sum_\ell \phi^\ell_{pc}\Delta^\ell_{pc} + \phi^\ell_{pn}\Delta^\ell_{pn} + \phi^\ell_{cp}\Delta^\ell_{cp} + \phi^\ell_{cn}\Delta^\ell_{cn} + \phi^\ell_{np}\Delta^\ell_{np} + \phi^\ell_{nc}\Delta^\ell_{nc}
\end{aligned}
$$

where λ_j is the classic 2SLS weight from Imbens and Angrist (1994) arising from instrumenting for incarceration (treatment $p2$) with judge indicators, and

$$
\begin{aligned}
\phi^\ell_{st} &= (\widetilde{\phi}^\ell_{st} - \widetilde{\phi}^\ell_{ts})\mathbb{1}[\widetilde{\phi}^\ell_{st} - \widetilde{\phi}^\ell_{ts} > 0]\\
\widetilde{\phi}^\ell_{st} &= \sum_{j=1}^{J} \frac{\lambda_j}{E[D_p|z_j] - E[D_p|z_{j-1}]} \mathbb{1}[D_{\ell j}{=}s, D_{\ell,j-1}{=}t]\pi_\ell
\end{aligned}
$$

This decomposition of $\beta^{2\text{SLS}}_{incar}$ exploits the fact that for each pair of treatments s and t, the 2SLS-weighted judge assignment induces a compliance group either from s to n or vice versa. By construction, ϕ^ℓ_{mn} is always weakly positive and when it is strictly positive represents the weight on the treatment effect for individuals who move from treatment t to s when they are assigned to the j^{th} rather than the $j-1^{\text{th}}$ most severe judge.

We then define the average treatment effect for individuals induced from treatment n to m as

$$
\Delta_{st} = \frac{\sum_\ell \phi^\ell_{st}\Delta^\ell_{st}}{\sum_\ell \phi^\ell_{st}} \qquad \text{(A2)}
$$

and the weight as $\phi_{st} = \sum_{\ell} \phi_{st}^{\ell}$. This lets us rewrite the 2SLS estimand, which is discussed in the main text as (8), as

$$\beta_{incar}^{\text{2SLS}} = \phi_{pc}\Delta_{pc} + \phi_{pn}\Delta_{pn} + \phi_{cp}\Delta_{cp} + \phi_{np}\Delta_{np} + \phi_{cn}\Delta_{cn} + \phi_{nc}\Delta_{nc} \tag{A3}$$

A2 Instrumenting for multiple treatments at once

In this section, we discuss the interpretation of the 2SLS estimand when the researcher simultaneously instruments for both conviction and incarceration. We show that the estimands reflect a combination of the effects of different treatments and for different complier groups, defying a causal interpretation even under restrictive patterns of compliance. We conclude that the results of these regressions should be interpreted as causal effects only under the assumptions of constant effects.

We consider the following 2SLS specification:

$$Y_i = \beta_0 + \beta_c \mathbb{1}[D_i{=}c] + \beta_p \mathbb{1}[D_i{=}p] + \varepsilon_i \tag{A4}$$

$$\mathbb{1}[D_i{=}c] = \alpha_j^c + e_i \tag{A5}$$

$$\mathbb{1}[D_i{=}p] = \alpha_j^p + u_i \tag{A6}$$

where β_c and β_p are the coefficients of interest and α represents the judge indicators used as the instruments.

By Frisch-Waugh-Lovell, these estimands can be decomposed as

$$\beta_c = \frac{\frac{\text{Cov}(Y,P_c)}{\text{Var}(P_c)} - \frac{\text{Cov}(P_c,P_p)}{\text{Var}(P_c)}\frac{\text{Cov}(Y,P_p)}{\text{Var}(P_p)}}{1-\rho_{cp}^2}$$

$$\beta_p = \frac{\frac{\text{Cov}(Y,P_p)}{\text{Var}(P_p)} - \frac{\text{Cov}(P_c,P_p)}{\text{Var}(P_p)}\frac{\text{Cov}(Y,P_c)}{\text{Var}(P_c)}}{1-\rho_{cp}^2}$$

where $P_d = E[D{=}d|Z=j]$ for each judge j and $\rho_{cp} \equiv \text{Cov}(P_c, P_p)/\sqrt{\text{Var}(P_c)\text{Var}(P_p)}$ is the correlation between P_c and P_p.

The expression for β_c reveals that the coefficient on the conviction dummy in (A4) is equal to the coefficient from a 2SLS regression of the outcome on instrumented conviction, minus the coefficient from a 2SLS regression of the outcome on instrumented incarceration multiplied by the effect of judge-instrumented conviction on incarceration, all rescaled by a term involving the correlation between the two treatment propensities.

Analogously to the single-treatment case in Appendix A1, these expressions can be decomposed into a weighted combination of treatment effects for the compliers corresponding to compliance group ℓ across each of the three treatments. In particular, it is easy to use the arguments in Appendix A1 to show that

$$\beta_c = \sum_{\ell} \phi_{21}^{\ell c}\Delta_{21}^{\ell} + \phi_{20}^{\ell c}\Delta_{20}^{\ell} + \phi_{12}^{\ell c}\Delta_{12}^{\ell} + \phi_{10}^{\ell c}\Delta_{10}^{\ell} + \phi_{02}^{\ell c}\Delta_{02}^{\ell} + \phi_{01}^{\ell c}\Delta_{01}^{\ell} \tag{A7}$$

where

$$\phi^{\ell c}_{mn} = \max\Big(0, (\widetilde{\phi}^{\ell c}_{mn} - \psi_c \widetilde{\phi}^{\ell p}_{mn}) - (\widetilde{\phi}^{\ell c}_{nm} - \psi_c \widetilde{\phi}^{\ell p}_{nm})\Big)$$

$$\widetilde{\phi}^{\ell d}_{mn} = \sum_{j_d=1}^{J} \frac{\lambda^d_j}{E[D_d|z_{j_d}] - E[D_d|z_{j_d-1}]} \mathbb{1}[D_{\ell j_d}{=}m, D_{\ell, j_d-1}{=}n]\pi_\ell$$

and where λ^d_j is the classic 2SLS weight from Imbens and Angrist (1994) arising from instrumenting for being in treatment d with judge indicators. $\psi_c = \frac{\text{Cov}(P_c,P_p)}{\text{Var}(P_c)}$ is the coefficient from a regression of judges' p treatment share on their c treatment share. We sub-index the judges indices with d to denote that the ordering is treatment-specific. The expression for β_p is the same as that for β_c, but using $\phi^{\ell p}$ weights.

The weights in (A7) depend on the underlying compliance patterns and are positive by construction, but do not necessarily sum to one. This means that it is in general not possible to interpret β_c and β_p as positively-weighted combinations of treatment effects, since they will in general include compliers moving in opposite directions across treatments.

These estimands remain uninterpretable even under restrictive substitution patterns. In Table A6, we calculate the weights on the different compliance groups under a single-index model of treatment defined in SI. We consider a case with three equally-likely judges who have incarceration and conviction thresholds of $(g_2(z), g_1(z)) \in \{(0.25, 0.30), (0.35, 0.75), (0.40, 0.85)\}$. This generates 5 different compliance groups ($u \in \{(0.25, 0.30], (0.30, 0.35], (0.35, 0, 40], (0.40, 0.74], (0.75, 0.85]\}$). Note that these judges satisfy the stringent OM condition; moving from judge 0 to 1 moves defendants from n to either c and p, while moving from judge 1 to 2 moves some defendants from c to p, and other defendants from n to c.

Nonetheless, the table reveals that the coefficients from (A4) will reflect a combination of treatment effects for different compliance groups. The coefficient on β_p reflects moves from conviction into incarceration (with a weight of 1.4), but also moves from conviction to dismissal (two different compliance groups, for a total weight of 1.6). It will also reflect (c, n), (n, p), and (c, p) effects. Under treatment effect heterogeneity, therefore, β_p will not correspond to an effect of incarceration in any meaningful sense.

In fact, it is difficult to find any situation in which the 2SLS coefficients will represent margin-specific causal effects. Consider the thresholds $(g_2(z), g_1(z)) \in \{(0.0, 0.0), (0.0, 1.0), (0.5, 1.0)\}$, which implies judge 0 convicts no one, judge 1 convicts everyone and incarcerates no one, and judge 2 incarcerates half and convicts the rest. Despite there being judge-pair comparisons such as judges 1 to 0 and judges 2 to 0 that each identify margin-specific effects of conviction and incarceration and therefore satisfies the so-called unordered partial monotonicity (UPM) assumption in Mountjoy (2022), the weights on the incarceration term ϕ^{lp}_{mn} have equal weight of 0.5 on (c, n), (p, c), and (p, n). As such, the weights do not reflect comparisons of margin-specific causal treatment effects and additionally include compliance for treatment effects, such as dismissal to conviction effects, that do not reflect the treatment of interest. Therefore, even in the most propitious conditions, the coefficients mix effects across non-target compliance groups.

There are two important selections. First, as discussed in Bhuller and Sigstad (2022), if the judge propensities for treatments c and p are linear in each other, then 2SLS continues to deliver a positive-weighted combination of treatment effects under the single-index model. This condition is testable, and if it happens to hold in any particular dataset this may be a viable alternative.

Second, under constant treatment effects (i.e., if $\Delta_{mn}^{\ell} = \Delta_{mn}$ for all m, n, and ℓ), then it is easy to show that $\beta_c = \Delta_{cn}$ and $\beta_p = \Delta_{pn}$. We conclude that unless the researcher can be confident that there is no heterogeneity in treatment effects across compliance groups or if the judge propensities are linear in each other and a single-index model of treatment assignment is appropriate, simultaneously instrumenting for multiple treatments is unlikely to make the coefficients in (A4) interpretable.

A3 Proofs of index representation propositions

A3.1 Proof of Proposition 2

Following Vyltacil (2002), we have that Assumption JM can be equivalently written as

$$
\begin{aligned}
1\{D_{cp}(z) = 1\} &= 1\{U_1 \leq g_1(z)\} \ , \\
1\{D_p(z) = 1\} &= 1\{U_2 \leq g_2(z)\} \ ,
\end{aligned}
$$

where $U_1, U_2 \sim U[0,1]$, and $g_1(z) = 1 - P(D(z) = n)$ and $g_2(z) = P(D(z) = p)$. As

$$D(z) = p1\{D_p(z) = 1,\ D_{cp}(z) = 1\} + c1\{D_p(z) = 0,\ D_{cp}(z) = 1\} + n1\{D_{cp}(z) = 0\} \ ,$$

the threshold crossing equation in (9) then directly follows. Next, to obtain the restriction in (10) and (11), observe that since logically $P[D_{cp}(z) = 0,\ D_p(z) = 1] = 0$, it follows that

$$P[U_1 > g_1(z),\ U_2 \leq g_2(z)] = 0 \ . \tag{A8}$$

Moreover, since $P(D(z) = n) + P(D(z) = c) + P(D(z) = p) = 1$ and $P[D(z) = c] = P[U_1 \leq g_1(z),\ U_2 > g_2(z)]$, we have

$$P[U_1 \leq g_1(z),\ U_2 > g_2(z)] = g_1(z) - g_2(z) \tag{A9}$$

This completes the proof.

A3.2 Proof of Corollary 2.1

Since $\mathcal{Z}$ is such that for every $t \in [0,1]$ there exists $z \in \mathcal{Z}$ such that $g_1(z) = g_2(z) = t$, it directly follows from (10) and (11) that

$$
\begin{aligned}
P[U_1 > t,\ U_2 \leq t] &= 0 \ , \\
P[U_1 \leq t,\ U_2 > t] &= 0,
\end{aligned}
$$

for all $t \in [0,1]$. This implies $P(U_1 = U_2) = 1$.

A3.3 Proof of Proposition 3

The proof is identical to the first part of that of Proposition 2.

A4 Testable implications of model assumptions across judges

We show testable implications from the single-index (SI) and latent monotonicity (LM) assumptions on binary judge comparisons. Consider any defendant characteristic X, although we will assume it ranges between 0 and 1 for simplicity.[1] We use expressions for Wald estimands over XD_p instrumenting for D_p, which provides information on complier characteristics of incarcerated defendants, and over XD_n, instrumenting for D_n, providing information on dismissed defendants. By examining treatment-specific characteristics of defendants, we isolate treatment margins that are restricted by the underlying model assumptions and provide bounds for the estimands. We compare judges $Z = 1$ and $Z = 0$, and let $p_{ij} = P[D(1) = i, D(0) = j]$.

The Wald estimands can be rewritten as:

$$\begin{aligned}\gamma_p &= \frac{E[XD_p|Z=1] - E[XD_p|Z=0]}{E[D_p|Z=1] - E[D_p|Z=0]}\\ &= \frac{E[X|D(1)=p, D(0)=n]p_{pn} + E[X|D(1)=p, D(0)=c]p_{pc}}{p_{pn} + p_{pc} - p_{np} - p_{cp}}\\ &+ \frac{E[-X|D(1)=n, D(0)=p]p_{np} + E[-X|D(1)=c, D(0)=p]p_{cp}}{p_{pn} + p_{pc} - p_{np} - p_{cp}}\end{aligned}$$

and

$$\begin{aligned}\gamma_n &= \frac{E[XD_n|Z=1] - E[XD_n|Z=0]}{E[D_n|Z=1] - E[D_n|Z=0]}\\ &= \frac{E[X|D(1)=n, D(0)=p]p_{np} + E[X|D(1)=n, D(0)=c]p_{nc}}{p_{np} + p_{nc} - p_{pn} - p_{cn}}\\ &+ \frac{E[-X|D(1)=p, D(0)=n]p_{pn} + E[-X|D(1)=c, D(0)=n]p_{cn}}{p_{np} + p_{nc} - p_{pn} - p_{cn}}\end{aligned}$$

The single-index assumption implies

$$\begin{aligned}0 \leq \gamma_p \leq 1\\ 0 \leq \gamma_n \leq 1.\end{aligned}$$

To see the relation for γ_p, assume $P[D_p|Z = 1] \geq P[D_p|Z = 0]$. This implies $g_p(1) \geq g_p(0)$, which is the sole cutoff threshold for D_p, and consequently $p_{np} = p_{cp} = 0$, i.e. there are no defiers moving out of treatment p. Finally X is bounded between 0 and 1, so the numerator is bounded $[0, p_{pn} + p_{pc}]$, establishing the result. For the relation for γ_d, assume $P[D_n|Z = 1] \geq P[D_n|Z = 0]$. This implies $g_n(1) \geq g_n(0)$, and consequently $p_{pn} = p_{cn} = 0$.

The latent monotonicity assumption implies

$$\begin{aligned}-\infty \leq \gamma_p \leq \infty\\ 0 \leq \gamma_n \leq 1.\end{aligned}$$

[1] If X has a wider range, the bounds on the Wald estimands we derive simply scale by the size of this range. In addition, we note that X could be endogenous to the treatment but is not required.

This model defines the same cutoffs g_c as in the single-index case, and consequently $p_{pn} = p_{cn} = 0$ for the D_n moments, leading to the relation on γ_n. For γ_p, there is no restriction on compliance types. To see that γ_p is unbounded consider a condition in which the first stage denominator is 0, but the numerator is positive or negative. This is possible given that $p_{np} + p_{nc} - p_{pn} - p_{cn} \in [-1, 1]$ and the weighted average in the numerator is not restricted based on the first stage coefficients. Intuitively, multiple judge thresholds control entry into treatment p under LM, so knowing the share does not restrict the potential for two-way flows into and out of this treatment across judges.

The above conditions can be tested using the methods developed in Frandsen, Lefgren and Leslie (2023). This method was designed to test single treatment IV assumptions for exclusion and monotonicity violations. It can instead be used to test the implications of the single-index or latent monotonicity assumptions on the appropriately defined outcome and treatment moments discussed here as the between-judge slope conditions are similar.[2] A clear benefit of this approach is that the inference procedure is designed to account for estimation error in the judge propensities.

Table A5 presents results from the semi-parametric "fit" test across court-year cells. The χ^2 test statistics are aggregated across court-year cells to provide a joint test, since the test statistics and associated DOFs can be summed under the assumption of independence. We run tests with D_n and D_p interacted with two covariates, any past charge and any future charge over the 5 years post-filing. Columns (1) and (3) show that we cannot reject the test on the D_n moments for either covariate ($p = 1$). This moment condition is the only test of the latent montonicity assumption, and hence the data provides some support for the underlying assumption.

We also test the D_p moments, which uniquely is implied by the single-index model. Column (2) shows we reject the test for the variable of any past charge, $\chi^2(DOF) = 1349(1208), p = 0.005$, while column (4) shows we cannot reject for any future charges $\chi^2(DOF) = 1204(1208), p = 0.53$. Rejecting the test on the D_n moments indicates that the data appear to be inconsistent with the implications of the single-index assumption. Together, these model implications and associated tests provide evidence against the single-index and instead provide some support that data is consistent with the latent monotonicity assumption.

[2] This test may be conservative as the Frandsen, Lefgren and Leslie (2023) test is designed to bound the between-judge differences to be -1 and 1 for an outcome with a 0 to 1 range.

A5 Identification in Humphries et al. (2023)

Another approach to identification of multiple treatment effects in examiner designs is in Humphries et al. (2023). Their method shares some similarities with our approach, but differs in other important respects. In contrast to the standard examiner assignment design—and in contrast to our approach—their method relies on the existence of a special class of covariates in addition to the examiner identity for identification. In this section we elaborate on these differences.

They use a multinomial choice model of judge decision making, where

$$D(z) = \underset{d \in \{n,c,p\}}{\text{argmax}}\ U_{idzx} \tag{A10}$$

$$U_{idzx} = \begin{cases} 0 & \text{if } \ d = n, \\ \beta_{dx} - g_d(z) + \varepsilon_{idzx} & \text{if } \ d \in \{c,p\} \ , \end{cases} \tag{A11}$$

and where the distribution of ε_{idzx} is known up to a finite-sized parameter vector of length α. β_{dx} represents the effect of covariates on decisions, and $g_d(z)$ represents the judge-specific thresholds.[3]

To understand how identification works in their setting, it is informative to simplify their model to consider a case with no covariates and where judges are unconditionally randomly assigned. This is usually considered the best-case situation in examiner designs; the classic treatments of instrumental variable models (e.g. Imbens and Angrist, 1994) assume unconditional random assignment of instruments and do not even include a discussion of covariates. We can write the distribution of U as

$$U_{idzx} = \begin{cases} 0 & \text{if } \ d = n, \\ -g_d(z) + \varepsilon_{idzx} & \text{if } \ d \in \{c,p\} \ , \end{cases} \tag{A12}$$

This choice equation gives rise to the following relationship between the judge thresholds g and the judge-specific choice propensities:

$$\begin{aligned} P[D=n|Z=z] &= \int_{-\infty}^{g_c(z)} \int_{-\infty}^{g_p(z)} f(u_1,u_2)\ du_2\ du_1 \\ P[D=c|Z=z] &= \int_{g_c(z)}^{\infty} \int_{-\infty}^{g_p(z)-g_c(z)+u_1} f(u_1,u_2)\ du_2\ du_1 \\ P[D=p|Z=z] &= \int_{g_p(z)}^{\infty} \int_{-\infty}^{g_c(z)-g_p(z)+u_2} f(u_1,u_2)\ du_1\ du_2 \end{aligned} \tag{A13}$$

where F is the distribution of ε. However, from this representation it is clear that unless this

[3] Their specification on page 33 does not include β_{dx}, but they allow the intercepts and variances of the random effect component of ε_{idzx} to vary by district and year. (A11) merely takes the intercept outside of ε_{idzx} and relabels it β_{dx}. Thus, although our specification is isomorphic to theirs, we view it as allowing a clearer understanding of the underlying identification assumptions.

distribution is known, g cannot be identified. Indeed, as in our setting, each distribution F implies a different mapping between the observed choice propensities and thresholds g that rationalize them. This can be seen more concretely in Figure A2, which takes F as a normal copula with correlation ρ. For a judge that assigns 30, 20, and 50% of defendants to treatments n, c, and p, respectively, we plot $(g_1(z), g_2(z))$ for each $\rho \in [0, 0.2, 0.4, 0.6, 0.8, 1]$. The judge thresholds g vary considerably with ρ, illustrating that in the standard no-covariates case, a multinomial choice model cannot be identified without prior knowledge of F.

Another way to see this is to simply count up the number of parameters and unknowns. For each judge, there are two moments (because the choices are mutually exclusive, any two of the choice propensities imply the third). If there are a parameters that govern the distribution of F, then there are $2|Z| + a$ parameters but only $2|Z|$ moments. Thus, even in the simple case where F has a single parameter, g is not identified.

This poses a threat to the interpretation of any such model, because F is precisely the parameter that governs the substitution behaviour across judges as well as the the existence and size of the compliance groups. One option—which we pursue in this paper—is to accept that the first stage is only partially identified, and that in turn many of the parameters of interest are also only partially identified. Humphries et al. (2023) instead rely on the existence of an additional set of regressors that allow them to point-identify g under additional restrictive assumptions on the data-generating process.[4]

These regressors are represented by β_{dx} in (A11). Since they are separable from the judge effects g, they are implicitly assumed to shift the distribution of the unobservables ε by the same amount for each judge. In combination with a parametric assumption on the distribution of ε, this allows identification of all the parameters of the model. In particular, if each judge appears in each mutually exclusive and exhaustive x cell, there are $2|Z||X|$ moments $P[D=d|Z=z, X=x]$ but only $2|X| + 2(|Z| - 1) + a$ parameters. Thus, the existence of these additional variables allows identification even in a setting with two judges, two values of X, and $\alpha \leq 2$.[5]

In this light, identification in Humphries et al. (2023) should be understood as arising *because* of the existence of these additional variables, rather than X being a simple nuisance parameter.[6] The selection of X is therefore a key design choice in the separable model, and should be something that theory suggests would shift the distribution of the individual-level unobservables ε without affecting the judges' thresholds in index space. It is unclear how their chosen X—court district and year—satisfies this condition, since these types of variables are typically used as conditioning variables in examiner designs, rather than as an integral part of identification.[7] Indeed, if the judges were unconditionally randomly assigned, which is usually

[4] In an Appendix, they estimate a version of (A12) where ε is distributed as a standard logit. Since this distribution has no parameters, in this case the thresholds g are identified without the existence of additional variables for identification, but at the cost of assuming that the compliance patterns are known ex ante.

[5] In this case, there are 8 moments and $2 \times 2 + 2 \times (2 - 1) + a$ parameters.

[6] This can be seen particularly clearly in the case where judges are only observed within a single x cell, or equivalently that the judge effects are nonseparable in X. Then, there are again only $2|Z|$ moments but $2|Z| + a$ parameters, and the model is unidentified.

[7] A partial list of papers that condition on court or time variables but use only the variation arising from the examiner effects for identification includes Arnold, Dobbie and Yang (2018), Arteaga (2021), Bhuller et al.

the best-case situation, each $P[D{=}d|Z{=}z, X{=}x]$ moment would approach $P[D{=}d|Z{=}z]$ as the sample size grew. This means that there would be fewer moments than parameters, leading to the unusual scenario of non-identification only under unconditional random assignment.

More broadly, it does not appear that *any* choice of X would satisfy the conditions required for nonparametric identification of F and g. Humphries et al. (2023) appeal to Berry and Haile (2023)—which develops a very general model in the context of demand estimation—to argue that their model is identified. However, Assumption 5 in Berry and Haile (2023) requires that the instrument Z be continuous. The judge indicators, in contrast, are discrete and hence these results do not apply.

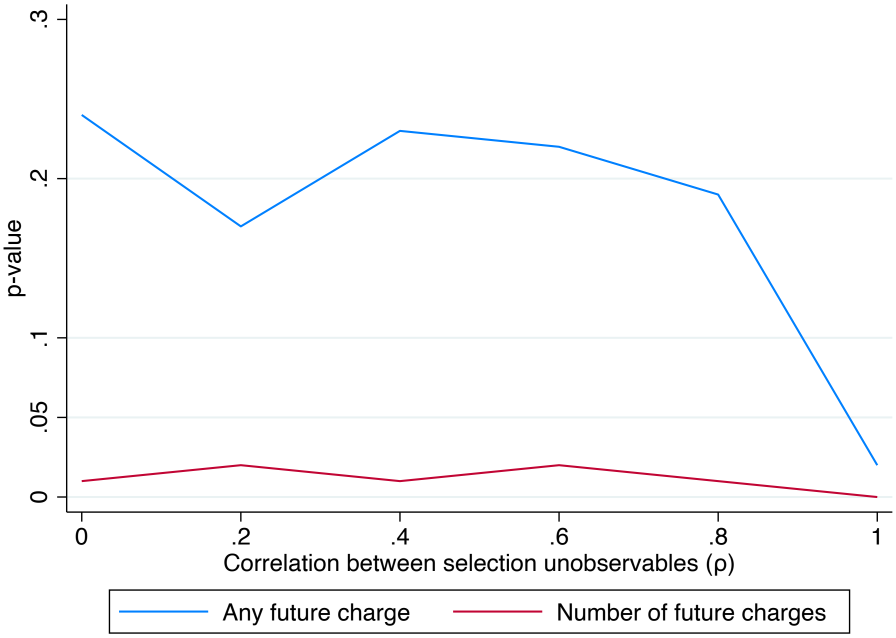
A final issue in Humphries et al. (2023) arises in identifying the MTRs after one has identified the judge thresholds. For identification they rely on Mountjoy (2022), who shows how to use local variation in one margin while controlling for the other. However, Mountjoy assumes that the instruments are continuous. Continuous instruments play an important role in allowing the researcher to difference out the "comparable compliers" across margin-specific shifts. The judge instruments are discrete—meaning, for example, that judges will likely never exactly match each other on one margin but differ on another—but it is unclear how to extend the comparable compliers assumption to this more complicated setting. This problem appears to be nontrivial: Mountjoy (2022) notes that "identifying margin-specific LATEs with discrete instruments would thus require additional assumptions about homogeneity in potential outcomes across these different complier groups," but does not conduct any further analysis. Humphries et al. (2023) provides no guidance on this important point.

To summarize, Humphries et al. (2023) differs from our approach in two important ways. First, they rely on a combination of separability and the existence of special additional covariates beyond the judge indicators for identification of the first stage. Second, they implicitly extend Mountjoy (2022) to discrete instruments, but do not clarify the assumptions required for this approach to be identified or demonstrate that they are satisfied. It thus departs from the standard practice of using judge instruments alone for identification, and from known identification results.

(2020), Dobbie, Goldin and Yang (2018), Doyle et al. (2015), Green and Winik (2010), Gross and Baron (2022), Kling (2006) and Maestas, Mullen and Strand (2013). Even approaches that interact instruments with covariates for power (e.g., Mueller-Smith (2015)) are in principle identified without the interactions. We were unable to find any papers that rely on the presence of judge-characteristic interactions for identification.

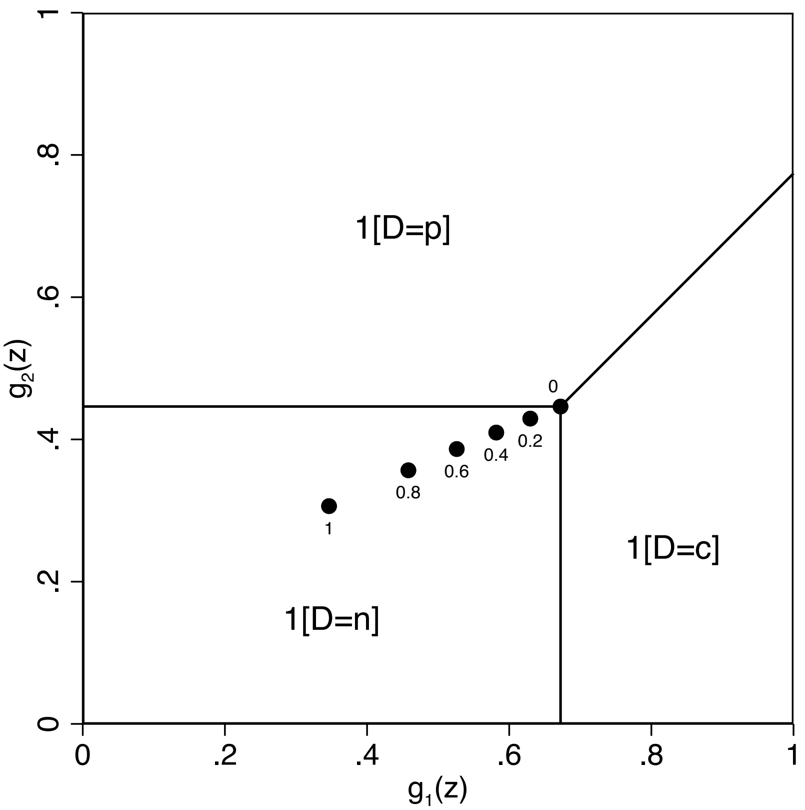
A6 Appendix Figures

Figure A1: *p*-values for null of matching 2SLS estimate of effect of incarceration, by ρ



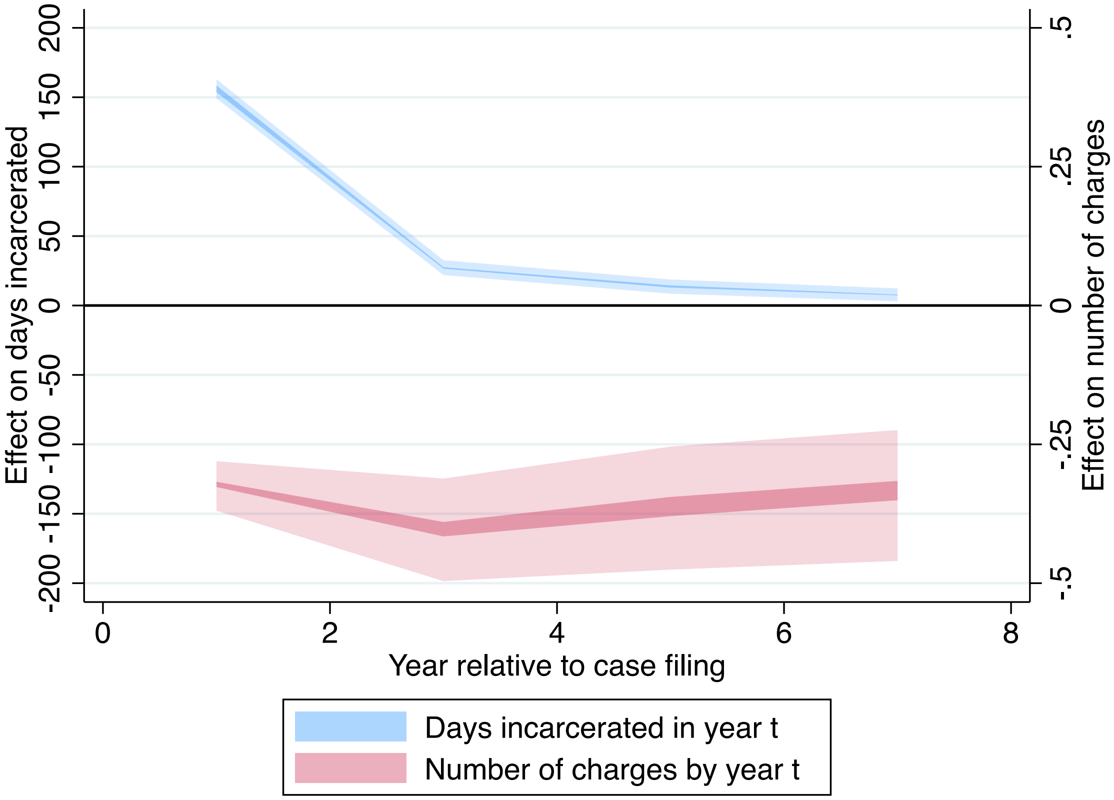
This figure shows the estimated *p*-values for a test of the null hypothesis that the structural model recovers the 2SLS coefficient from a regression of the given outcome on incarceration, where incarceration is instrumented by judge indicators. Distribution of 2SLS coefficients under null estimated via a boostrap with 200 draws.

Figure A2: Judge thresholds g in multinomial choice model when U is distributed as a normal copula with varying correlation ρ



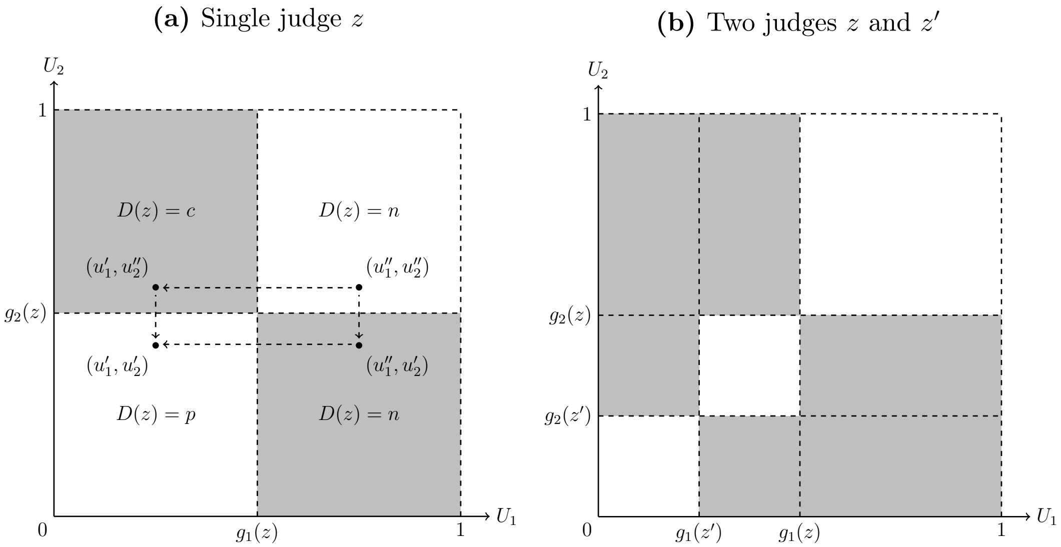
For $\rho \in [0, 0.2, 0.4, 0.6, 0.8, 1]$, this graph plots $(g_1(z), g_2(z))$ for a judge that assigns 30, 20, and 50% of defendants to treatments n, c, and p, respectively.

Figure A3: Effect of incarceration relative to conviction on incapacitation and future charges for felony cases



This figure shows estimated effects of incarceration relative to conviction only (Δ_{pc}) on the number of days spent incarcerated until year t and number of charges up until year t, for $t = 1, 3, 5$, and 7 years after case filing. The sample is restricted to felony cases. The darker areas denote the range of estimates arising from choice models with selection parameters $\rho \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$. The lighter areas denote the edge of 95% confidence intervals for the endpoints.

Figure A4: Inadmissable regions space of unobservables under JM



(a) Single judge z

(b) Two judges z and z'

This figure displays different regions implied by Proposition 2 for an example with a single judge z with $g_1(z) = g_2(z)$ and how the inadmissible region (corresponding to the shaded gray areas) increases with an additional judge z' with thresholds equal to half of those of z. Note that we drop the conditioning on $x \in \mathcal{X}$ for convenience.

A7 Appendix Tables

Table A1: Defendant characteristics, by court

	Felony	Misdemeanor
Black	0.62	0.52
	(0.54)	(0.60)
Age	32.23	31.47
	(11.03)	(11.98)
Any previous charge	0.56	0.47
	(0.44)	(0.49)
Property crime	0.39	0.20
	(0.46)	(0.37)
Drug crime	0.33	0.23
	(0.46)	(0.40)
Violent crime	0.16	0.22
	(0.32)	(0.40)
Convicted	0.87	0.52
	(0.29)	(0.44)
Incarcerated	0.29	0.10
	(0.42)	(0.27)
Sentence length	187.38	3.83
	(546.37)	(26.96)
Observations	363,032	355,222

Sample is all defendants in Cuyahoga, Hamilton and Franklin county courts. Standard deviations in parentheses.

Table A2: Effects of conviction and incarceration on future criminal justice outcomes

				Never prev. convicted	
	All	Fel.	Misd.	Fel.	Misd.
	(1)	(2)	(3)	(4)	(5)
Panel A: Number of charges over next 5 years					
Incarceration rel. to conviction (Δ_{pc})	[0.040, 0.252]	[0.069, 0.133]	[0.017, 0.319]	[-0.143, 0.037]	[-0.013, 0.224]
	(-0.117, 0.359)	(-0.175, 0.431)	(-0.176, 0.447)	(-0.347, 0.344)	(-0.237, 0.375)
Incarceration rel. to not guilty (Δ_{pn})	[0.486, 2.897]	[0.196, 0.541]	[0.593, 4.023]	[-4.315, -0.076]	[1.002, 4.570]
	(0.016, 5.564)	(-4.146, 4.538)	(0.001, 7.659)	(-9.852, 1.223)	(0.387, 8.224)
Conviction rel. to not guilty (Δ_{cn})	[-0.189, 0.017]	[-0.222, 0.027]	[-0.194, 0.014]	[-0.103, 0.058]	[0.026, 0.131]
	(-0.310, 0.162)	(-0.512, 0.232)	(-0.361, 0.224)	(-0.352, 0.260)	(-0.113, 0.330)
Panel B: Number of convictions over next 5 years					
Incarceration rel. to conviction (Δ_{pc})	[-0.028, 0.193]	[0.145, 0.347]	[-0.131, 0.210]	[0.060, 0.304]	[-0.111, 0.198]
	(-0.216, 0.332)	(-0.028, 0.698)	(-0.414, 0.379)	(-0.149, 0.624)	(-0.424, 0.400)
Incarceration rel. to not guilty (Δ_{pn})	[0.698, 4.048]	[-0.628, 0.385]	[0.964, 5.988]	[-3.607, -0.010]	[1.428, 6.335]
	(-0.232, 8.329)	(-6.224, 4.969)	(0.060, 11.413)	(-9.320, 2.106)	(0.350, 11.134)
Conviction rel. to not guilty (Δ_{cn})	[0.054, 0.254]	[-0.247, 0.054]	[0.105, 0.329]	[-0.030, 0.140]	[0.246, 0.342]
	(-0.103, 0.449)	(-0.561, 0.286)	(-0.092, 0.591)	(-0.279, 0.326)	(0.054, 0.586)
Weight on $c \to p$ effect	[0.334, 0.444]	[0.316, 0.438]	[0.341, 0.446]	[0.278, 0.358]	[0.268, 0.339]
Weight on $n \to p$ effect	[0.000, 0.237]	[0.000, 0.366]	[0.000, 0.190]	[0.000, 0.292]	[0.000, 0.160]
Weight on $n \to c$ effect	[0.768, 1.005]	[0.639, 1.005]	[0.815, 1.005]	[0.714, 1.006]	[0.845, 1.006]

This table reports treatment effects of conviction and incarceration, aggregated using the weights from a 2SLS regression with conviction as the treatment and judge dummies as the instruments. MTRs are approximated by a second-degree polynomial in u_1 and u_2 as specified in Section 5.5.

Table A3: Effects of conviction and incarceration on future criminal justice outcomes

				Never prev. convicted	
	All	Fel.	Misd.	Fel.	Misd.
	(1)	(2)	(3)	(4)	(5)
Panel A: Number of charges over next 7 years					
Incarceration rel. to conviction (Δ_{pc})	[-0.214, -0.174]	[-0.350, -0.316]	[0.010, 0.091]	[-0.298, -0.206]	[-0.015, 0.033]
	(-0.305, -0.106)	(-0.453, -0.219)	(-0.111, 0.219)	(-0.437, -0.103)	(-0.175, 0.178)
Incarceration rel. to not guilty (Δ_{pn})	[0.165, 2.171]	[-3.095, -0.467]	[0.724, 4.544]	[-2.400, 0.299]	[0.992, 3.554]
	(-1.159, 5.500)	(-9.270, 3.080)	(0.206, 7.228)	(-11.078, 6.279)	(0.224, 6.883)
Conviction rel. to not guilty (Δ_{cn})	[-0.176, 0.069]	[-0.375, -0.130]	[-0.072, 0.174]	[-0.227, 0.063]	[0.243, 0.414]
	(-0.338, 0.252)	(-0.867, 0.191)	(-0.279, 0.425)	(-0.474, 0.377)	(-0.003, 0.718)
Panel B: Number of convictions over next 7 years					
Incarceration rel. to conviction (Δ_{pc})	[-0.295, -0.250]	[-0.405, -0.363]	[-0.131, -0.036]	[-0.303, -0.177]	[-0.038, 0.038]
	(-0.427, -0.132)	(-0.549, -0.225)	(-0.366, 0.194)	(-0.475, -0.059)	(-0.261, 0.256)
Incarceration rel. to not guilty (Δ_{pn})	[0.176, 2.805]	[-4.765, -0.751]	[0.942, 6.201]	[0.258, 1.368]	[1.253, 4.677]
	(-1.663, 7.274)	(-14.072, 4.542)	(-0.107, 11.086)	(-10.274, 10.790)	(-0.137, 9.490)
Conviction rel. to not guilty (Δ_{cn})	[0.085, 0.382]	[-0.468, -0.020]	[0.195, 0.598]	[-0.115, 0.166]	[0.532, 0.785]
	(-0.207, 0.653)	(-1.120, 0.355)	(-0.142, 1.005)	(-0.565, 0.543)	(0.205, 1.141)
Weight on $c \to p$ effect	[0.977, 1.000]	[0.990, 1.000]	[0.949, 1.000]	[0.989, 1.000]	[0.967, 1.000]
Weight on $n \to p$ effect	[0.000, 0.054]	[0.000, 0.038]	[0.000, 0.084]	[0.000, 0.044]	[0.000, 0.073]
Weight on $n \to c$ effect	[0.062, 0.084]	[0.030, 0.045]	[0.122, 0.158]	[0.040, 0.057]	[0.117, 0.144]

This table reports treatment effects of conviction and incarceration, aggregated using the weights from a 2SLS regression with incarceration as the treatment and judge dummies as the instruments. MTRs are approximated by a second-degree polynomial in u_1 and u_2 as specified in Section 5.5.

Table A4: Policy effects after 7 years

	Change in treatment shares			Effects on outcomes	
	n	c	p	N. charges	N. conv.
Panel A: Felony defendants					
Incarceration leniency ($g_2 \downarrow$)	0.000	[0.009, 0.010]	[-0.010, -0.009]	[0.329, 0.757]	[0.298, 0.834]
				(0.200, 1.057)	(0.106, 1.259)
Conviction leniency ($g_1 \downarrow$)	0.010	[-0.010, -0.007]	[-0.003, 0.000]	[-0.093, 0.219]	[-0.043, 0.155]
				(-0.396, 0.448)	(-0.412, 0.476)
No incarceration ($g_2 = 0$)	0.000	0.292	-0.292	[0.077, 0.308]	[0.127, 0.321]
				(-0.055, 0.374)	(-0.029, 0.489)
No conviction ($g_1 = 0$)	0.874	-0.582	-0.292	[-11.842, -11.779]	[-4.454, -4.370]
				(-24.501, 0.856)	(-22.968, 14.062)
Panel B: Misdemeanor defendants					
Incarceration leniency ($g_2 \downarrow$)	0.000	[0.005, 0.010]	[-0.010, -0.005]	[-0.246, 0.205]	[-0.283, 0.368]
				(-0.397, 0.428)	(-0.523, 0.753)
Conviction leniency ($g_1 \downarrow$)	0.010	[-0.010, -0.008]	[-0.002, 0.000]	[-0.151, 0.185]	[-0.481, -0.003]
				(-0.366, 0.369)	(-0.776, 0.265)
No incarceration ($g_2 = 0$)	0.000	0.107	-0.107	[-0.035, 0.039]	[-0.051, 0.049]
				(-0.056, 0.070)	(-0.085, 0.103)
No conviction ($g_1 = 0$)	0.539	-0.433	-0.107	[-0.775, -0.766]	[-1.259, -1.245]
				(-1.442, -0.090)	(-2.098, -0.392)
Panel C: Never-convicted misdemeanor defendants					
Incarceration leniency ($g_2 \downarrow$)	0.000	[0.005, 0.010]	[-0.010, -0.005]	[-0.179, 0.144]	[-0.279, 0.210]
				(-0.355, 0.399)	(-0.538, 0.599)
Conviction leniency ($g_1 \downarrow$)	0.010	[-0.010, -0.008]	[-0.002, 0.000]	[-0.295, -0.061]	[-0.598, -0.297]
				(-0.541, 0.128)	(-0.943, -0.028)
No incarceration ($g_2 = 0$)	0.000	0.089	-0.090	[-0.019, 0.015]	[-0.027, 0.023]
				(-0.036, 0.041)	(-0.053, 0.059)
No conviction ($g_1 = 0$)	0.536	-0.446	-0.090	[-0.704, -0.693]	[-1.189, -1.173]
				(-1.346, -0.044)	(-1.947, -0.406)

Table reports the effects of a number of policy changes on recidivism. We analyze marginal changes, which shift judges' thresholds g by 0.01, as well as global changes. The change in treatment shares is the change from the given policy. The change in outcomes is rescaled by the number of defendants whos' treatment is affected by the policy change for the marginal changes to assist in readability. Bounds are in square brackets and the outer edges of 95% confidence intervals in parentheses.

Table A5: Using the Frandsen-Lefgren-Leslie test to test model assumptions

	(1) Any past charge $\times D_n$	(2) Any past charge $\times D_p$	(3) Any charges $\times D_n$	(4) Any charges $\times D_p$
χ^2	828	1337	850	1201
Deg. of freedom	1208	1208	1208	1208
p	1	.00532	1	.55
Observations	638,684	638,684	638,684	638,684

This table displays results from the semi-parametric Frandsen test to adjudicate between choice models. Outcomes of the form XD_n apply the Frandsen-Lefgren-Leslie test with judges instrumenting for D_n, and XD_p does so for D_p. The table reports results from the fit component of the Frandsen-Lefgren-Leslie test, applied with 3 knots and done separately across court-year cells. The chi-square test statistics and degrees of freedom are aggregated across cells to test the joint condition.

Table A6: Weights on compliance groups in two-treatment 2SLS

	$\phi^{\ell c}_{mn}$						$\phi^{\ell p}_{mn}$					
Group ℓ	(c,n)	(p,n)	(p,c)	(n,c)	(n,p)	(c,p)	(c,n)	(p,n)	(p,c)	(n,c)	(n,p)	(c,p)
1	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2
2	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0
3	0.2	0.0	0.0	0.0	0.0	0.4	0.0	0.0	1.4	0.2	0.0	0.0
4	1.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.4	0.0	0.0
5	0.0	0.0	0.0	0.8	0.0	0.0	2.8	0.0	0.0	0.0	0.0	0.0

This table reports the weights for the different compliance groups generated in a regression where both incarceration and conviction are instrumented for with judge assignment. We assume that treatment is determined by the single-index model defined in SI, and that there are three equally-likely judges who have incarceration and conviction thresholds of $(g_2(z), g_1(z)) \in \{(0.25, 0.30), (0.35, 0.75), (0.40, 0.85)\}$. This generates five compliance groups, with $u \in \{(0.25, 0.30], (0.30, 0.35], (0.35, 0, 40], (0.40, 0.74], (0.75, 0.85]\}$. $\phi^{\ell c}_{mn}$ is the weight of the $m \to n$ effect for group ℓ in the coefficient on treatment c; $\phi^{\ell p}_{mn}$ is the analogous weight for the p coefficient.