Prison, Probation, and the Community

Sam Kang (University of St. Gallen)*

August 29, 2024

Abstract

Probation plays a key role in managing the prison population and provides large potential fiscal and social gains, but little is known about the broader externalities to the community of a probationed offender. This paper investigates these questions by leveraging the random assignment of criminal cases to judges in order to estimate the causal effect courtroom punishments on the future criminal behavior of both the defendant and their community. I derive new identification results for two-stage least squares that illustrate the risk of omitted treatment biases and show how these biases can be addressed using measures of judge preferences as instruments. In my empirical analysis, I find that both prison and probation sentences lead to reductions in recidivism, but that these punishments operate through alternative mechanisms. Prison acts to prevent crime through incapacitation while probation acts to reduce crime through increased policing and re-conviction. At the community level, I find that greater exposure to criminal offenders leads to higher crime rates by other community members, which suggests a multiplier effect associated with prison sentences but not probation sentences.

*This draft: June 2024. Author contact: samueljong.kang@unisg.ch. I thank Beatrix Eugster and Scott Cunningham for their invaluable support and advice throughout the development of this project. I also thank Giovanni Mastrobuoni and Brigham Frandsen for their helpful comments and feedback.

1 Introduction

The US carceral population stands out as one of the largest amongst developed countries. At its peak in 2007, the US incarceration rate reached 506 inmates per 100,000 US residents, a number over 260% larger than the contemporaneous prison rate across Europe (Guerino et al., 2011; Walmsley, 2015). The social and economic costs brought by incarceration to this level have been well-documented in the literature and have sparked ongoing debates for decarceration. In turn, the past decade has witnessed new efforts across all levels of government to reduce the inmate population. This has been met with a dramatic decline in the carceral population of approximately 20% since 2012 (Carson and Kluckow, 2023). Probation plays a key role in managing incarceration rates, as the probation population measures at approximately 250% the size of the prison population (Kaeble, 2023).

Probation provides offenders with an opportunity to avoid incarceration by instead serving a supervisory sentence within their community. The potential benefits are clear. For one, the fiscal savings of probation are expectedly large as estimates suggest that over 80% of probationed offenders avoid incarceration (Kaeble, 2023). Two, probation can mitigate negative effects associated with incarceration as past research has shown that imprisonment contributes to the development of criminal capital, labor market exclusion, and the formation of criminal peer networks (Bayer et al., 2009; Drago and Galbiati, 2012; Aizer and Doyle, 2015; Bhuller et al., 2018a; Mueller-Smith and T. Schnepel, 2020). These effects of incarceration collectively contribute to high recidivism rates and more severe future criminal behavior (Mueller-Smith, 2015).

However, a largely unexplored aspect is how probation might influence the local community of an offender and it is ex-ante unclear whether probation is socially improving for members of the offender's community and household. Theoretically, reductions in incarceration will be met with higher steady state crime rates due to lower incapacitation and deterrent effects (Becker, 1968; Kessler and Levitt, 1999; Drago et al., 2009; Buonanno and Raphael, 2013; Barbarino and Mastrobuoni, 2014; Bhuller et al., 2018b). It would be expected then that a higher reliance on probation will increase the risk of victimization and generally increase exposure to harmful or risky behaviors and influences (Norris et al., 2021; Arteaga, 2023). On the other hand, the incarceration of a household member can both destabilize household dynamics and finances while also bringing negative emotional and developmental impacts for household members (Dobbie et al., 2018a; Bhuller et al., 2018a).

In this paper, I study how incarcerative and probationary sentences shape the future criminal behavior of offenders and how exposure to offenders in turn influences the local criminal activity of the household and community. To causally identify effects of sentencing, I implement an instrumental variable estimator with a "judge fixed effect" design that leverages the randomized assignment of judges to criminal defendants.

In the first step of my analysis, I derive new identification results for the two-stage least squares (2SLS) estimator of multiple endogenously assigned and continuously-valued treatments. My results formalize the conditions required for 2SLS to be interpretted as a weighted LATE in the presence of arbitrary first and second stage heterogeneity and can be viewed as extensions of the monotonicity and exclusion restriction assumptions from lower dimensional settings. These results bridge the recent literature studying the performance of 2SLS with presence of multiple discrete treatments (Bhuller and Sigstad, 2023; Frandsen et al., 2023; Humphries et al., 2024) with the literature on continuously-valued treatments (Angrist et al., 2000). My results also more broadly relate to the recent literature on regression estimators for continuously valued treatments in the presence of heterogeneous treatment effects (Goodman-Bacon, 2021; Callaway et al., 2024).

Importantly, I show that in the context of court room sentencing decisions, 2SLS estimators are only valid when the researcher instruments for both the type of punishment received by an offender (e.g., prison or probation) and the length of the sentence and formalizes the arguments made in (Mueller-Smith, 2015). Past research sutdying spillovers of judge sentencing decisions have generally not adopted this approach and relied on alternative approaches to controlling for "omitted treatment biases" (Arteaga, 2023; Bhuller et al., 2018a; Dobbie et al., 2018b).

My empirical findings show that both prison and probation reduce future crime by offenders, but through different mechanisms. Probation prevents crime by increasing the policing on probationed offenders, thereby increasing re-conviction rates for those who remain criminally active. As a result, there is a short-term increase in convictions the probation sentences that fades as re-offenders are convicted again. The effect on prison differs in that it reduces crime through incapacitation. That is, it prevents crime in the short-run by holding offenders in prison, but I do not find that the response to be released from prison is influenced by the duration of the prison sentence. In studying community responses, I find only exposure to ex-convicts increases the community crime rate. In particular, crime rates from those living at the same address as the defendant rise the longer that a previously incarcerated defendant is in community.

2 Institutional Setting

My analysis focuses on the city of Milwaukee, Wisconsin over the years 2006 to 2018. All criminal cases originating from within the city are also handled by the city's courthouse. The courthouse features six criminal divisions that handle different classifications of crimes. These six divisions are as follows: general felonies, felonies of sexual assault and homocide, drug felonies, gun felonies, general misdemeanors, and misdemeanors of domestic violence. Cases in the four specialty divisions may be selectively assigned to judges. For this reason, I decide to only focus on cases from the general felony and general misdemeanor divisions where judges are randomly assigned to cases.

2.1 Case Assignment

In the Milwaukee courthouse's general felony and general misdemeanor divisions, criminal cases are assigned to judges through a multistep process. This sequence is described by the flowchart in Figure 1. First, cases are assigned to one of the six criminal divisions based upon the classification of the charges. After this, cases are assigned to a judge within the assigned division.

Judges may serve in only one division at a time and may serve within the criminal court for a maximum of four years before they are rotated into a different branch of the courthouse (e.g., the family courts). Thus, during a judge's tenure within a particular division, they will only be assigned to cases that have been first assigned to that division.

Major court rotations are dictated annually by the courthouse's chief judge at the start of August, in which approximately 25% of judges are rotated across court branches. Minor rotations are also made over the course of the year when, for example, a judge is promoted to a higher circuit court, retires, or takes an extended leave of absence. In such cases, judges may be rotated

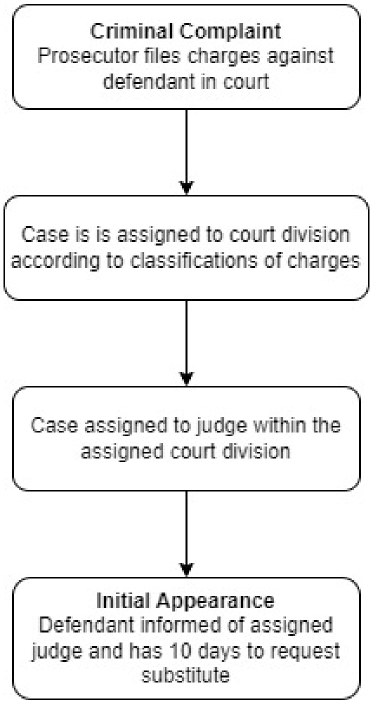

Criminal Complaint
Prosecutor files charges against defendant in court
Case is is assigned to court division according to classifications of charges
Case assigned to judge within the assigned court division
Initial Appearance
Defendant informed of assigned judge and has 10 days to request substitute

Figure 1: Milwaukee Judicial Assignment Process

across divisions in order to address vacancies. I refer to periods between any judicial rotation within a division (major or minor) as a “division rotation”.

Once a judge has been assigned to a case, the defendant is informed and has ten days to request a substitute judge. Upon request for a substitute judge, the case will again be randomly assigned to a different judge within the division. This assignment is final and a defendant may not request a second substitution. Substitution requests can induce endogeneity in the assignment process by allowing defendants to select away from unfavorable matches. To address this, my analysis will always focus on the effect of the originally assigned judge regardless of whether a substitution is made.

2.2 Sentencing

The assigned judge oversees the trial and, upon an independent jury’s determination that the defendant is guilty, the judge will decide on the defendant’s sentence. The Wisconsin penal system features discretionary sentencing in which the judge decides the punishment of a guilty defendant from a range of possible punishments defined by the penal codes. In general, criminal sentences may contain both an incarceration sentence and a fine. Wisconsin also features truth-in-sentencing laws that generally require a defendant to serve the entirety of any incarceration sentence given by the judge at the time of sentencing.

A judge may alternatively place a guilty defendant on probation. In doing so, the judge simultaneously determines both the probation sentence and the “stayed” prison sentence. The stayed sentence defines an alternative prison sentence that the defendant receives only if they violate the terms of their probation.

In my analysis, I consider a criminal sentence to be comprised of the following dimensions:

whether or not the defendant is incarcerated, whether or not the defendant is placed on probation, whether or not the defendant is fined[1], and the lengths of any prison, probation, or stayed sentence.

3 Data

To conduct my analysis, I combine Wisconsin state court records and the Milwaukee Master Property Assessment Files (MPROP) for the years 2004 to 2023. During this period, there are 148,248 publically visible criminal cases within the Milwaukee criminal court. My sample is defined as all guilty defendants in the general felony or general misdemeanor court divisions that committed a crime after January 1, 2005 and whose trial ended before December 31, 2018, for which there are 56,252 cases[2]. To match to the MPROP files, I additionally restrict my sample to those defendant whose recorded address is within the city of Milwaukee. This excludes 21,065 additional cases of defendants who are not residents of Milwaukee city. Cases for which the assigned judge oversees less than 10 defendants outside of the defendant's residential zip code are also excluded to ensure that the judge fixed effect instrument is sufficiently predictive of judge preferences. This excludes an additional 5,838 cases. Lastly, I additionally exclude 1,062 cases for whom the defendant is guilty but is neither incarcerated nor placed on probation. This selection provides easier interpretation of results at the cost of a possible sample selection bias. Accordingly, I will also assess the robustness of my results when including these defendants. My final sample consists of 28,287 trials[3].

3.1 Court Records

In Wisconsin, state court records are made publically available online through the Consolidated Court Automation Program (CCAP). For criminal and eviction trials, records are posted for all defendants who lose their trial, while family court records are posted for all cases. I collected all publically available criminal, family, and eviction court records in Milwaukee from 2004 to 2023 directly through CCAP[4].

CCAP criminal court records report the defendant's legal name, race, date of birth, sex, and address as well as the assigned judge, charged offenses, determination of guilt or innocence by charge, and final sentences. Records for guilty defendants remain posted for a minimum period of 20 years and so I observe sentences for all guilty offenders during my collected sample.

Family courts cover trials for interventions by the child protective services, claims for alimony and child support, divorce proceedings, and restraining orders. Family records report the name, age, and address of both the defendant and the plaintiff and, unless the case is dismissed, remains posted for a minimum of 40 years. I observe all family trials that are carried out during my sample period.

[1] Fine amounts are not uniformally reported or measured in the Wisconsin court records and are prone to significantly large variation in the reported quantities and so I leave this out of my analysis.

[2] My sample begins in 2005 to allow for a sufficient pre-trial period over which pre-trial variables can be measured. My sample ends in December 2018 because the covid pandemic introduced emergency measures for court procedures. As well, this allows for an appropriate post-trial period to study evictions as the national eviction moratorium began in March 2020.

[3] In practice, the sample consists of 19 "singletons" that are dropped from the regressions. Thus, I often present summary statistics and results with the effective sample that omits these singletons.

[4] Data collection of court records was carried out between June 2023 to January 2024.

Eviction court records include the self-reported name of the defendant, the address from which the defendant is evicted from, and the final judgement of the case. All records include details on dates of trial proceedings and events, such as the dates of the initial filing and final judgement. Eviction records remain posted for a minimum of 20 years when the defendant loses the trial (i.e., an eviction actually occurs), and so I also observe all executed evictions during my sample period.

3.2 MPROP

The MPROP files provide an annual census of all properties in the city of Milwaukee and is collected by the Milwaukee Assessor's Office. It includes information on the property owner and descriptive characteristics of the property, such as the size of the lot or floorspace in the building, number of units in the building, the use of the property, and the assessed value of the property. Each property may be uniquelly identified by either the address or a property tax ID assigned by the city.

3.3 Data Linking

Criminal court records provide accurate identifying information of defendants, including name and date of birth. This allows me to track the criminal trials against a defendant over time by linking criminal records on the name and date of birth of the defendant. To link property-level measures back to the defendant, I first match all court records to the listed address' property ID using the city's online property search look-up[5]. This allows me to accurately link all address-level data in the city. I then study the response of a defendant's residence by linking cases matched to the same property ID as the defendant but committed by a different person.

3.4 Summary Statistics

Table 1 presents summary statistics on the defendants within my sample. Panel A presents statistics for defendants, Panel B characterizes the crimes committed, while Panel C presents statistics about their residences. The average defendant is approximately 30 years old (26 at the median) with over 84% of defendants being male. The large majority of defendants are black, comprising over 79% of the sample. White defendants make up nearly 18% with the remainder of defendants being of other racial background. Approximately 13% of offenders were observed committing a crime within the previous year before their trial. Divorce and paternity claims were experienced by only 4% and 5% of the sample within the year prior to the defendant's offense while over 16% of defendants were evicted in the preceding year, signalling a high eviction rate within the sample. The sample is nearly evenly split between felony and misdemeanor offenders, with 53% of defendants being on trial for a felony. The large share of defendants commit non-violent crimes, with only approximately 13% committing violent crimes.

Defendants tend to live on properties with multiple residential units (7.4 units on average and 2 units at the median). The average unit within the residence of defendant is a 2.4 bedroom unit with approximately 1,100 square feet. The average value of a unit on the property is valued at approximately $54,000 on average and $43,000 at the median.

[5] The look-up may be found at https://itmdapps.milwaukee.gov/MyMilwaukeeHome/indexSidebarNew.jsp#.

Table 1: Summary Statistics of Defendants

	Mean (1)	Median (2)	SD (3)	N (4)
Panel A: Defendant Characteristics				
Age of defendant	29.63	26.17	11.04	28,324
Male defendant	0.8450	1.00	0.3619	28,324
White defendant	0.1785	0	0.3829	28,324
Black defendant	0.7873	1.00	0.4092	28,324
Other race defendant	0.0342	0	0.1818	28,324
Committed crime in past year	0.1331	0	0.3397	28,324
Divorced in past year	0.0431	0	0.2030	28,324
Paternity claim in past year	0.0475	0	0.2127	28,324
Evicted in past year	0.1627	0	0.3691	28,324
Panel B: Case Characteristics				
Felony	0.5252	1.00	0.4994	28,324
Property crime	0.3122	0	0.4634	28,324
Violent crime	0.1254	0	0.3311	28,324
Other crime	0.5625	1.00	0.4961	28,324
Panel C: Property Characteristics				
Units in property	7.40	2.00	24.79	28,324
Mean area of unit (sqft)	1,075	1,056	583	28,324
Mean value of unit (USD)	54,501	43,100	123,716	28,324
Mean # bedrooms	2.36	2.00	1.15	28,324
Owner occupied	0.0356	0	0.1852	28,324
Crimes in previous year	0.0802	0	1.81	28,324

Notes: Summary statistics of means, medians, and standard deviations for defendants in my primary sample. Panel A presents statistics for defendant characteristics, Panel B presents statistics for case details, and Panel C presents characteristics describing defendant's residences.

4 Modelling Judge Sentencing Decisions

Here I introduce a model on judicial sentencing decisions. The model first considers a potential criminal who decides whether or not to commit crime and faces a certain probability of punishment. If apprehended, the judge decides on a punishment for the criminal in consideration of how it may effect future criminality. The primary purpose of the model is to provide a general framework for understanding how the identifying assumptions of 2SLS estimators induce a certain structure on judge biases and preferences. This model generally encompasses the threshold crossing framework adopted in related studies, but also allows for a broader set of judge sentencing patterns. However, given the importance of threshold crossing models in the literature, I discuss this as a special case below and in Appendix D.

Defendant's Problem

Consider a setting in which individuals live for two time periods. In each period, individuals choose a bundle of M crimes denoted by $C \in \mathbb{R}_+^M$. A criminal i receives benefits according to the function B_i, but face a risk of punishment given by q_0. The punishment is defined by the $2K$-dimensional vector W. The punishment W is comprised of variables $D_k \in \{0, 1\}$ and $T_k \in \mathbb{R}_+^K$ where, for example, D_{prison} indicates whether the offender is sentenced to prison and T_{prison} indicates the length of the prison sentence. For a given punishment w, the defendant receives disutility according to the function S_i.

$$B_i : \mathbb{R}_+^M \to \mathbb{R}_+$$
$$q_0 : \mathbb{R}_+^M \to [0, 1]$$
$$S_i : \mathbb{R}_+^{2K} \to \mathbb{R}_+$$

The function q_0 is assumed known but W is ex-ante unknown. Instead, criminals know the expected value of W for a given bundle c, denoted as $\mu_w(c) := \mathbb{E}[W|c]$. Individual i's chosen bundle, denoted by C_{1i}, maximizes the following equation

$$\max_c \pi_i(c) = B_i(c) - q_0(c)\mathbb{E}[s(c)].$$

The above setup induces a probability distribution over C_1, given by $\Phi_1(c \mid q_0, \mu_w(c))$:

$$\Phi_1\big(c \mid q_0, \mu_w(c)\big) := \mathbb{P}\Big[C_1 = c \mid q_0(c), \mu_w(c)\Big].$$

Judge Sentencing Decisions

Judges j observe period 1 criminals i with characteristics v_i and decide on a punishment W. A punishment has three effects. One, the punishment may increase the probability of being punished for future crimes. Two, the punishment informs the defendant about future potential punishments. Three, the punishment can change a criminal's future incentives for crime by remapping the functions B and S. This induces a new probability distribution over criminal behavior, $\Phi_2(c_2)$. For a given punishment w, these effects are summarized below:

$$\begin{aligned} q_0(c) &\to q_w(c) \geq q_0(c) \\ \mu_w(c) &\to w \\ \Phi_1(c) &\to \Phi_2(c|q_w, w) \end{aligned}$$

Judges receive disutility if the criminal commits crime in period 2 according to the M-dimensional vector of parameters $\alpha_j < \mathbf{0}$ and also receive disutility from administering punishments, given by the function $R_j : \mathbb{R}^K_+ \to \mathbb{R}$ which satisfies $R'_j < 0$ and $R''_j < 0$. However, judges do not know B_i or S_i but instead observe V_i. Furthermore, judges hold biased beliefs regarding the distribution Φ_2. Specifically, judge j believes that

$$\Phi_j(c|v_i, q_w, w) = \Phi_2(c|v_i, q_w, w).$$

Judges choose a punishment W to maximize expected disutility subject to a constraint that $w \in supp(W)$. The constraint captures that legal codes may define sets of possible punishments that judges may administer. Thus, the judge's maximization problem is

$$\max_w \mathbb{E}[U_{i,j}] = R_j(w) + \alpha_j \int_{\mathbb{R}^K_+} c \; d\Phi_j(c|q_w, w, v_i) \;\; s.t. \;\; w \in supp(W).$$

Denote judge j's maximizing choice for the punishment by w_j. Then we can generally express the set of potential assignments of punishments to criminal i by the function

$$h(j, c_{1i}, v_i) = w_j$$

Special Case: Threshold Crossing Model

Threshold crossing models are encompassed within the general framework described above. In the context of discrete treatment settings, threshold crossing models provide a powerful framework to estimate marginal treatment effects following the approach of Heckman and Vytlacil (2005). In the continuous treatment setting, threshold crossing models are still useful for developing a grounded understanding of estimated treatment effects. For this reason, I include a brief discussion relating my results to the above setup here, but for a more complete discussion direct readers to related studies like Arteaga (2023); Bhuller and Sigstad (2023); Chyn et al. (2024); Humphries et al. (2024).

A simple variation of a threshold crossing model can be obtained by assuming that judges hold common preferences over offenses α_j and common beliefs Φ_j, that Φ_j is rank invariant in v with respect to w, and that preferences over punishments R_j are commonly ordered. This is summarized below:

$$\begin{array}{rcll} \Phi_j(c|q_w, w, v_i) & = & \Phi_{j'}(c|q_w, w, v_i) & \forall \, (j, j') \\ \alpha_j & = & \alpha & \forall \, j \\ \Phi_j(c|q_w, w, v_i) > \Phi_j(c|q_w, w, v'_i) & \Leftrightarrow & \Phi_j(c|q_{w'}, w', v_i) > \Phi_j(c|q_{w'}, w', v'_i) & \forall \, (w, w', v_i, v'_i) \\ R_j(w) > R_j(w') & \Leftrightarrow & R_{j'}(w) > R_{j'}(w') & \forall \, (j, j', w, w') \end{array}$$

The restrictions on Φ_j and α_j induces a single index for measuring the severity of defendants that is common to all judges and hence variation in sentencing only arises from differences in stringency. The intuition of the above model is that all judges share a common ordering of punishments W and a common ordering of defendants i in terms of their perceived severity. Furthermore, these orderings are independent of each other and defined across the support of v and so the ranking of an individual is invariant to the choice of W[6]. Hence, we can express an individual i's indexed ranking along a single index $\phi_i \in [0,1]$.

Estimating with Judge Fixed Effects

The model suggests that a defendant's sentence can be viewed as a function h that depends both on characteristics of the criminal, given by v_i and c_{1i}, and characteristics of the assigned judge j. And while sentences W are generally endogenously determined, the assignment of the judge can be viewed as inducing an exogenous sentencing shock. This shock is expressed as

$$\delta_i = h(j, c_{1i}, v_i) - \mathbb{E}[h(J, c_{1i}, v_i)|c_{1i}, v_i].$$

Borusyak and Hull (2023) show that (a variance-weighted) δ_i is an efficient and exogenous measure of sentencing shocks that controls for influence of the characteristics c_1 and v. That is, if the δ_i were observed (or equivalently $\mathbb{E}[h(J, c_{1i}, v_i)|c_{1i}, v_i]$), then one could estimate the effect of sentencing shocks directly with the following OLS regression

$$Y_i = \beta\delta_i + \gamma_r + u_i$$

where γ_r is a fixed effect for the randomization group (in my setting, this is the $Crime \times Division \times Cycle$) and u is a stochastic error term. However, for each i we only observe $W_i = h(j, c_{1i}, v_i)$ and hence $\mathbb{E}[h(J, c_{1i}, v_i)|c_{1i}, v_i]$ must be estimated. This presents an empirical challenge because neither v nor the functional form of h are known.

The existing "judge fixed effect" literature has generally approached this with two-stage least squares estimators in which the chosen instrument Z represents some prediction of $h(J, c_{1i}, v_i)$. Most commonly used is the judge's average punishment, which is computed as a "leave-out" sample average that is calculated by excluding the individual i. The first stage involves estimating the sentencing shock with a randomization group fixed effect as follows

$$W_i = \alpha Z_i + \gamma_r + v_i \tag{1}$$

where γ_r is a randomization group fixed effect and v_i is a stochastic error term. The estimate for δ_i is given by αZ_i, which represents the difference in the mean punishment for the assigned judge relative to all other judges in the randomization group. The second stage regression is given by

$$Y_i = \beta^{2SLS}\alpha Z_i + \gamma_r + u_i. \tag{2}$$

[6] As disussed in Humphries et al. (2024), plea deals give a situation where the independence of defendant and punishment orderings are likely violated. In the context of the above model, this occurs when the cost function R is allowed to depend on whether a defendant accepts a plea. For clarity, let the punishment be given as $(W, Plea)$. Then the distribution $(W, Plea = 0, V)$ is likely very different to $(W, Plea = 1, V)$ because some individuals will never be offered the plea and some will never accept. As a consequence, Φ_j under $Plea = 1$ is not defined for some v. This can lead to violations in monotonicity as weaker punishments are accepted through plea deals by selectively more severe defendants.

Recent literature has emphasized that interpretation of β^{2SLS} is not trivial (Bhuller and Sigstad, 2023; Blandhol et al., 2022; Humphries et al., 2024). In general, IV estimators rely on the assumptions that Z is exogenous from U, that Z is properly excluded from the second stage equation, and that Z is strongly correlated with W. When either the first or second stage responses are homogeneous, these assumptions are sufficient to ensure that β^{2SLS} can be interpretted as a LATE. However, assuming homogeneous first and second stage responses is a strong assumption that is unlikely to be satisfied. In reality, judges are likely to have heterogeneous preferences over defendant characteristics (first stage heterogeneity) and defendants are likely to have heterogeneous responses to punishments (second stage heterogeneity). The results contained in Section 5 and also those of Bhuller and Sigstad (2023) and Humphries et al. (2024) show that in this more general case, β^{2SLS} only holds causal interpretations under additional assumptions, which can be viewed as multivariate extensions of the more familiar univariate monotonicity and exclusion restriction assumptions.

Even under the assumptions covered in Section 5, β^{2SLS} will generally represent a weighted average of treatment effects across different compliers. The weights generally are inversely proportional to how sensitive a complier is to the instrument. In the threshold cross model, this would mean that defendants who require larger sentencing shocks to be moved into harsher punishments will receive larger weights. This creates interpretation challenges when the weight distribution is also correlated with treatment effects. In the criminal court setting, that is likely to be the case because defendant who receive large weights will be those who are less severe, for whom the treatment effect is likely to be different than those who are more severe.

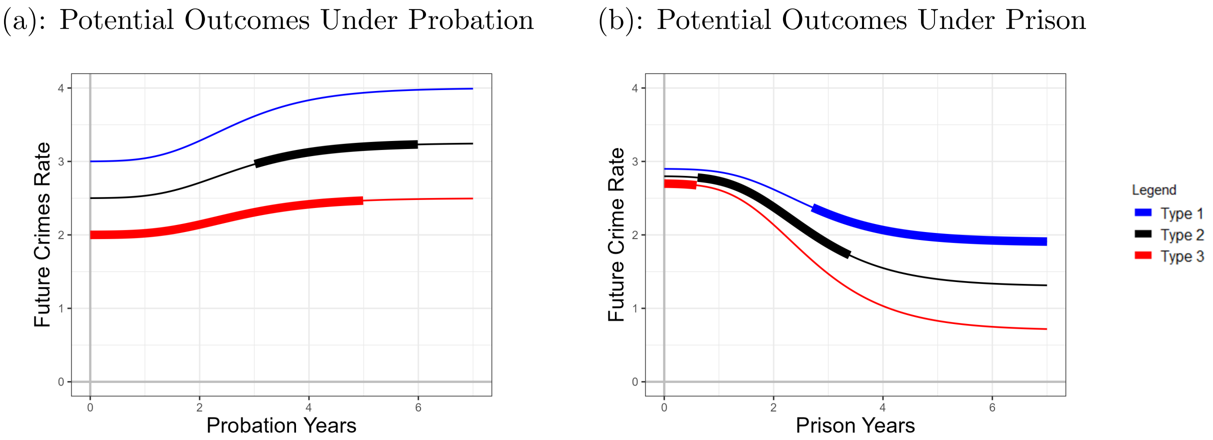
To address this, I will construct instruments that attempt to reduce the degree of unobserved first stage heterogeneity. Specifically, I will first use random forest estimators to predict the *potential* punishments with pre-trial defendant characteristics. In doing so, I will employ "honest" splitting rules, which Wager and Athey (2018) show that under appropriate assumptions, provide an unbiased estimate for the conditional mean function of the actual punishment assigned. I further reduce the risk of overfitting by training the random forests using samples that observation i and their zip code, so that the model predictions are uncorrelated with the actual assignment of the defendant or the assignment of local community members. By making predictions for all potential assignments that a defendant might receive across the set of potential assigned judges, a prediction for δ_i can then be made by taking the difference of the predicted punishment for the actually assigned judge from the average predicted punishment across the pool of potentially assigned judges and functions as a "re-centered" instrument in the style of Borusyak and Hull (2023). This instrument choice will have the benefit of reducing the degree of correlation between the complier weights and second-stage outcomes as the first stage heterogeneity should no longer be systematically correlated with the pre-trial covariates used in the random forest predictions.

5 Identification with Two Stage Least Squares

In this section, I provide identification results for the 2SLS estimator when the second stage endogenous variables W are possibly continuous. These results apply broadly to any 2SLS estimator with multiple continuous treatments as the setup allows for arbitrary heterogeneity in the both the first and second stages and the only restriction placed on the treatments is that they are assumed to be mutually exclusive and exhaustive. The instruments may generally take on any form but are assumed to satisfy the standard assumptions of exogeneity, exclusion, and rank, described below. These results are to my knowledge the first identification results for 2SLS estimators of multiple continuously valued endogenous treatments under arbitrary heterogeneity. The analysis may be viewed as an extension of the analysis of Bhuller and Sigstad (2023), Humphries et al. (2024), and Frandsen et al. (2023) to the continuous treatment setting or similarly as an extension of the analysis of Angrist et al. (2000) to the multiple treatment setting.

A key takeaway from my results is that if the instrument Z is correlated with all elements of W, then 2SLS is subject to an omitted treatment bias if any element of W is omitted. In the context of courtroom sentencing, this means that 2SLS will be biased and will not identify causal effects if for example the researcher only instruments for the extensive margin probability of being sentenced to prison. The intuition for this can be understood through the following example depicted in Figure 2. The lines in the figure depict the average potential outcome for three types of criminals, Type 1, Type 2, and Type 3. The bold regions on each line indicate the region on the support of W where that type might possibly be sentenced depending on the judge to which they are assigned. The thin regions on the lines indicate regions on the support of W where the respective type never receives the treatment under any judge assignment.

Figure 2: Average Potential Outcomes Example


(a): Potential Outcomes Under Probation

(b): Potential Outcomes Under Prison

Notes: Each figure plots the average potential outcome of a three response types under two alternative treatments. Response types are identified by the color of the line. Bold sections of the lines indicate regions on the support of the treatments at which the indicated response type might possibly receive the treatment.

The problem arises when one uses a research design that takes the assigned judge as an instrument for the punishment W, but defines W too narrowly. For example, a research design that only instruments for whether a defendant is sentenced to prison, but does not instrument for either the length of probation or prison sentences will violate the exclusion restriction. The goal of such a research design would be to estimate the effect of receiving a prison sentence relative to a probation sentence. Because Type 3 never receives a probation sentence under any judge assignment, valid estimates should not be influenced by the treatment status of Type 3. However, the judicial assignment does influence the sentence length for Type 3, and as a consequence judicial assignment instruments designed to predict the probability of being sentenced to prison will generally also be correlated with the length of the prison sentence. As a result, estimates are likely to capture the effect of both being sentenced to prison and of changing the length of the sentence at unrelated margins. Research designs that ignore this will confuse effects from having a harsher prison sentence with the effect of being incarcerated. This can more generally be viewed as a problem that arises when there is the simultaneous assignment of multiple treatments.

The analysis in the remainder of this section discusses how 2SLS designs decompose the effects of simultaneously assigned continuous treatments and under what conditions the estimand retains a causal interpretation. I begin by introducing the setup and notation. I then define the relevant treatment effect parameters that enter the 2SLS estimand. Finally I present identification results. In the main text I only present the setup and discuss the theoretical results while all proofs are presented in Appendix B.

5.1 Setup and Notation

Consider a vector W comprised of K mutually exclusive treatments which may be continuously valued. If treatment k is multivalued, then the vector W contains both the variable $D_k \in \{0,1\}$ indicating that the treatment k was received and the variable $T_k \in \mathbb{R}$ indicating the level received of the treatment. Then we may express W as

$$W = (D_1, \ldots, D_K, T_1, \ldots, T_K).$$

Let m denote the total number of elements in W. The outcome is denoted by $Y \in \mathbb{R}$ and the researcher has access to a vector of instruments $Z \in \mathbb{R}^n$ with $n \geq m$ that influences the assignment of W. Then I assume that Y and W can be represented as

$$\begin{aligned} Y &= g(W, U) \quad \text{(second stage)} \\ W &= h(Z, V) \quad \text{(first stage)} \end{aligned}$$

The function g is assumed to be continuously differentiable across the support of W. The terms U and V are first and second stage residual terms of unconstrained dimensionality. This setup allows for arbitrary and unconstrained heterogeneity across both the first and second stages responses. Potential outcomes are given by considering the function g at a fixed $W = w$ and $U = u$. "Response types" can also be defined according to the first stage residual V because after fixing $V = v$ the treatment assignment of W depends only on the instrument Z. In the threshold crossing framework presented in Section 4, response types are defined according to the common index of perceived defendant severity, Φ_j.

Lastly, because W is comprised of K mutually exclusive treatments, we may consider the potential outcomes conditional on receiving treatment k. In this case, the potential outcome depends only on the level of the treatment T_k and the second stage residual U. This is given by the function g^k defined as

$$g_i^k(T_k, U) \coloneqq g(D_1 = 0, \ldots, D_k = 1, \ldots, D_K = 0, T_1 = 0, \ldots, T_k, \ldots, T_K = 0, U).$$

Under the above setup, we may express the standard 2SLS assumptions as follows:

Assumption 1. *(Exogeneity)* $\{g(w,u), h(z,v), U, V\} \perp Z_i$

Assumption 2. *(Exclusion)* $h(z,v) = h(z',v)$ *implies* $g(h(z,v),u) = g(h(z',v),u)$

Assumption 3. *(Rank) $Cov(Z, W)$ has full rank*

Additionally, I assume that the treatment intensities T are non-negative and thus bounded from below by 0.

5.2 Treatment Effects

In the presence of multiple continuously valued treatments, there are several different notions of treatment effects that might be of interest to researchers and policymakers. I consider two particular treatment effects, which I denote the conditional "takeup effect" and the conditional average causal response on the treated (ACRT):

$$\begin{aligned}\beta_{k,v}^{takeup} &:= \frac{\mathbb{E}[g(W,U)|D_k = 1, v] - \mathbb{E}[g(W,U)\ |D_k = 0, v]}{\mathbb{E}[T_k|k, v]} \\ \beta_{k,v}^{ACRT}(t_k) &:= \mathbb{E}\Big[\frac{\partial}{\partial t} g^k(t, U)\big|_{t=t_k}\ |D_k = 1, v\Big]\end{aligned}$$

Each parameter is defined conditional on v and in reference to a particular treatment k. This means that the parameters may be interpretted as the group average effect for a specific response type, or, equivalently, as the average effect at a point on the severity index Φ_j. The takeup effect can be interpretted as the average change in the outcome under treatment k compared to all other treatment states, scaled by the expected dosage of treatment k. Thus, the takeup effect indicates how the potential outcome shifts on average when given treatment k. On the other hand, the average causal response indicates the average marginal response in the outcome when the treatment is marginally increased from a level t. The ACRT may be interpretted as the slope of the potential outcome curve under treatment k at the point t.

5.3 2SLS Estimand

Define the first stage prediction for W as $\widehat{W}$, where

$$\widehat{W} = Var(Z)^{-1}Cov(Z, W).$$

Then, under the above assumptions, the 2SLS estimand is identifed by the following result

Theorem 1. *Under Assumptions 1, 2, and 3, the 2SLS estimand is given by*

$$\beta^{2SLS} = \omega^{-1}\sum_{k=1}^{K}\int_0^\infty \mathbb{E}[\tau_{k,v}(t)\beta_{k,v}^{ACRT}(t)]dt + \omega^{-1}\sum_{k=1}^{K}\mathbb{E}\Big[\kappa_{k,v}\beta_{k,v}^{takeup}\Big]$$

where

$$\begin{aligned}\omega &:= \sum_{k=1}^{K}\int_0^\infty \mathbb{E}[\tau_{k,v}(t)]dt + \sum_{k=1}^{K}\mathbb{E}[\kappa_{k,v}] \\ \tau_{k,v}(t) &:= Var(\mathbb{1}\{T_k > t\}|v)\Big(\mathbb{E}[\widehat{W}|k, v, T_k > t] - \mathbb{E}[\widehat{W}|k, v, T_k \leq t]\Big) \\ \kappa_{k,v} &:= Var(D_k|v)\Big(\mathbb{E}[\widehat{W}|k, v] - \mathbb{E}[\widehat{W}]\Big)\mathbb{E}[T_k|k, v]\end{aligned}$$

The above result states that the 2SLS estimand can generally be interpretted as the expectation of weighted takeup effects and ACRT effects, where the expectation is taken across response types v. The effects for each response type are weighted by the τ and κ terms, which in expectation sum to ω (after integrating). Each parameter $\beta_{k,v}^{takeup}$ and $\beta_{k,v}^{ACRT}$ represents a local effect to response type v. Takeup parameters may be thought of as effects that are local to regions in $Supp(W)$ where discrete changes in the D_k are observed while ACRT parameters are local to regions in $Supp(W)$ where variation in T_k is observed.

However, Theorem 1 only states how the full parameter of coefficients in β^{2SLS} can be interpretted, but not the individual coefficients themselves. Theorem 2 identifies the individual coefficients by applying the Frisch-Waugh-Lovell Theorem. Specifically, first define the residualized first stage predictions, $\widetilde{W}$, as

$$\widetilde{W} := \Big(I_{2K} - \widehat{W}(\widehat{W}'\widehat{W})^{-1}\widehat{W}'\Big)\widehat{W}.$$

The residualized prediction for a particular element, $\widetilde{W}_k$, is obtained as the residual from regressing each element $\widehat{W}_k$ onto all other elements of $\widehat{W}$. Then by the Frisch-Waugh-Lovell Theorem, the coefficient β_k^{2SLS} is given by the regression of Y on $\widetilde{W}_k$. One may then apply Theorem 1 to the univariate case where $\widetilde{W}_k$ is the instrument for W_k, which gives the following result

Theorem 2. *Under Assumptions 1, 2, and 3, the 2SLS parameter* β_k^{2SLS} *is given by*

$$\beta_k^{2SLS} = \omega^{-1}\sum_{l=1}^{K}\int_0^\infty \mathbb{E}\Big[\beta_{l,v}^{ACRT}(t)\tau_{l,v}(t)\Big]dt + \omega^{-1}\sum_{l=1}^{K}\mathbb{E}[\kappa_{l,v}\beta_{l,v}^{takeup}].$$

where

$$\begin{aligned}\omega &:= \sum_{l=1}^{K}\int_0^\infty \mathbb{E}\Big[\tau_{l,v}(t)\Big]dt + \sum_{l=1}^{K}\mathbb{E}[\kappa_{l,v}]\\ \tau_{l,v}(t) &:= Var\big(\mathbb{1}\{T_l > t\}|v\big)\Big(\mathbb{E}[\widetilde{W}_k|l,v,T_l > t] - \mathbb{E}[\widetilde{W}_k|l,v,T_k < t]\Big)\\ \kappa_{k,v} &:= Var(D_l|v)\big(\mathbb{E}[\widetilde{W}_k|l,v] - \mathbb{E}[\widetilde{W}_k]\big)\mathbb{E}[T_k|k,v]\end{aligned}$$

The result of Theorem 2 generally mirrors that of Theorem 1, except that the weights now depends on $\widetilde{W}_k$ instead of $\widehat{W}$.

Identification of Weighted LATE

There are two important implications of Theorem 2. One, the effect of W_k for type v may enter the estimand β_k^{2SLS} with a negative weight. Two, the estimand β_k^{2SLS} for the effect of W_k may be contaminated by the effects of the other elements of W. The source of both of these problems is the first stage estimation. Negative weighting occurs when there is a systematic negative relationship between $\widetilde{W}_k$ with W_k. That is, for some response type v, $\widetilde{W}_k$ is on average decreasing as W_k increases. Cross-contamination occurs if after residualizing the first stage predictions with $\widetilde{W}$, there is still systematic correlation between $\widetilde{W}_k$ and the other dimensions of the treatment assignment. Under either condition, β_k^{2SLS} does not generally hold a causal interpretation. The remainder of the analysis focuses on determining necessary and sufficient conditions for ruling out these possibilities.

In order to ensure that β_k^{2SLS} may be interpretted as a weighted LATE, we require three components. The weights must sum to 1, the weights must be non-negative, and there must be no cross-contamination. The first property holds by construction. This can be seen by defining the set of weights for each response type v as

$$\omega_v = \omega^{-1} \sum_{k=1}^{K} \int_0^\infty \tau_{k,v}(t)dt + \omega^{-1} \sum_{k=1}^{K} \kappa_{k,v}.$$

Then by definition of ω we have that

$$\mathbb{E}[\omega_v] = 1.$$

The following two assumptions ensure that weights are non-negative and that there is no cross-contamination in effects.

Assumption 4. ***Average conditional monotonicity***
Discrete Case*: For all types v such that $Var(D_k|v) \neq 0$, $\mathbb{E}[T_k|k, v] \neq 0$, and $\mathbb{E}[\widetilde{D}_k|k, v] \neq \mathbb{E}[\widetilde{D}_k]$, the following holds*

$$sign(\mathbb{E}[\widetilde{D}_k|k, v] - \mathbb{E}[\widetilde{D}_k]) = sign(\mathbb{E}[T|k, v]).$$

Continuous Case*: For all points t in the support of T_k the following holds for any response type v satisfying $Var(\mathbb{1}\{T_k > t|v\}) \neq 0$,*

$$\mathbb{E}\Big[\widetilde{T}_k|k, v, T_k > t\Big] \geq \mathbb{E}\Big[\widetilde{T}_k|k, v, T_k \leq t\Big].$$

The average conditional monotonicity assumption states that the residualized prediction $\widehat{W}_k$ should on average be wekly increasing in the distribution of the actual W_k for *each* response type v. This requires that the first stage prediction on average is in the correct direction of the treatment level after removing linear correlation with other elements of the $\widehat{W}$.

Assumption 5. ***Conditional exclusion restriction***
Consider any element W_k of W. Then $\widetilde{W}_k$ satisfies the "conditional exclusion restriction" if the following mean independence condition holds:

$$\mathbb{E}[\widetilde{W}_k|v, W] = \mathbb{E}[\widetilde{W}_k|v, W_k]$$

The conditional exclusion restriction requires that the residualized first stage prediction $\widetilde{W}_k$ should be mean independent of all elements of W after conditioning on W_k for *each* response type v. Intuitively, this states that the residualized $\widetilde{W}_k$ first stage predictions should on average satisfy the exclusion restriction in a univariate IV estimator with W_k as the endogenous treatment.

The following result shows that the conditional monotonicity and conditional exclusion restriction assumptions are necessary and sufficient for β_k^{2SLS} to be interpretted as a weighted LATE.

Theorem 3. *Under Assumptions 1, 2, 3, the 2SLS coefficient on W_k, given by β_k^{2SLS}, represents a weighted LATE if and only if Assumptions 4, and 5 hold. The weighted averages are given by*

$$\beta_k^{2SLS} = \frac{\int_0^\infty \mathbb{E}[\tau_{k,v}(t)\beta_{k,v}^{ACRT}(t)]dt}{\int_0^\infty \mathbb{E}[\tau_{k,v}(t)]dt} \qquad (W_k = T_k)$$

$$\beta_k^{2SLS} = \frac{\mathbb{E}[\kappa_{k,v}\beta_{k,v}^{takeup}]}{\mathbb{E}[\kappa_{k,v}]} \qquad (W_k = D_k)$$

where the weights are positively valued for all v and given by

$$\begin{aligned}\tau_{k,v}(t) &:= Var\big(\mathbb{1}\{T_k > t\}|v\big)\Big(\mathbb{E}[\widetilde{W}_k|k,v,T_k > t] - \mathbb{E}[\widetilde{W}_k|k,v,T_k < t]\Big)\\ \kappa_{k,v} &:= Var(D_k|v)\big(\mathbb{E}[\widetilde{W}_k|k,v] - \mathbb{E}[\widetilde{W}_k]\big)\mathbb{E}[h(Z,V)|k,v]\end{aligned}$$

Theorem 3 states that under average conditional monotonicity and the conditional exclusion restriction, the 2SLS estimand can be interpretted as a weighted average of takeup effects (for discrete components D_k) or as a weighted average of ACRT parameters (for continuous components T_k).

Omitted Treatment Bias

Recent research has raised concerns about omitted treatment biases when some element(s) of W is excluded from the second stage equation (Mueller-Smith, 2015; Arteaga, 2023; Chyn et al., 2024). That is, let W_l be some element of W and W_{-l} represent the vector W when W_l is omitted. If Z is correlated with W_l, then omitting W_l from the second stage equation violates the exclusion restriction and 2SLS estimates will generally be biased. The following result formalizes this "omitted treatment bias".

Corollary 4. *Let $W = (D_1, \ldots, D_K, T_1, \ldots, T_k)$ and define W_{-l} as the vector including all elements of W except for some element W_l. Now suppose that Z satisfies Assumptions 1, 2, and 3 and is correlated with both W_{-l} and W_l.*

$$\beta^{2SLS} = \boldsymbol{\omega}^{-1}\sum_{k=1}^{K}\int_0^\infty \mathbb{E}[\tau_{k,v}(t)\beta^{ACRT}_{k,v}(t)]dt + \boldsymbol{\omega}^{-1}\sum_{k=1}^{K}\mathbb{E}\Big[\kappa_{k,v}\beta^{takeup}_{k,v}\Big]$$

where

$$\begin{aligned}\widehat{W}_{-l} &= Var(Z)^{-1}Cov(Z, W_{-l})\\ \boldsymbol{\omega} &:= \sum_{k=1}^{K}\int_0^\infty \mathbb{E}[\lambda_{k,v}(t)]dt + \sum_{k=1}^{K}\mathbb{E}[\kappa_{k,v}]\\ \tau_{k,v}(t) &:= Var(\mathbb{1}\{T_k > t\}|v)\Big(\mathbb{E}[\widehat{W}_{-l}|k,v,T_k > t] - \mathbb{E}[\widehat{W}_{-l}|k,v,T_k \le t]\Big)\\ \kappa_{k,v} &:= Var(D_k|v)\Big(\mathbb{E}[\widehat{W}_{-l}|k,v] - \mathbb{E}[\widehat{W}_{-l}]\Big)\mathbb{E}[T_k|k,v]\end{aligned}$$

The result mirrors that of Theorem 1, except that $\widehat{W}$ is replaced by $\widehat{W}_{-l}$. The intuition is that the effect of W remains unchanged when an element of W is omitted, but now the effect of the omitted treatment will be distributed across the parameters of the included treatments. A similar result holds in the form of Theorem 2 because the effect of l is still captured, but further refinements in the form of Theorem 3 will generally not hold. The consequence of this is that it is not possible to determine "where" the omitted treatment bias contaminates the coefficients of the included elements W_{-l}. It may be that the bias is entirely concentrated within an individual element of the treatment or it may be spread across all elements. This in general depends on the correlation of $\widetilde{W}_{-l}$ with W_l.

Comparing OLS and 2SLS

The results of Theorems 1, 2, and 3 relate to a broader literature that identifies the estimands of different regression estimators as a weighted average of heterogeneous treatment effects. Identification of the 2SLS estimand as a weighted average of heterogeneous treatment effects has been an on-going focus of the IV literature, but current research does not address the multi-dimensional continuous setting (for examples, see Angrist et al. (2000); Heckman and Vytlacil (2005); Bhuller and Sigstad (2023); Humphries et al. (2024); Frandsen et al. (2023); Blandhol et al. (2022)). Słoczyński (2022) provides results for OLS with a binary treatment that is exogeneous conditional on covariates and provides a comparison for the corresponding 2SLS estimator. Goodman-Bacon (2021) and Callaway et al. (2024) provide results for difference-in-differences designs in which the treatment might take a staggered adoption or feature continuous elements in assignment.

To better describe how my results relate to this broader literature, I provide a comparison of the 2SLS estimand to the OLS estimand for the effect of an endogenous treatment W for which the researcher has access to an exogenous instrument Z. This is provided in Appendix C. In general, the results show that both 2SLS and OLS estimators can be interpetted as weighted averages of heterogeneous treatment effects. However, OLS constructs weights across the joint distribution of (W, U) while 2SLS constructs weights across the joint distribution of (Z, V). The 2SLS estimand can be viewed as first marginalizing out the influence of U to estimate the group average treatment effect for response types v. These group average treatment effects are representative within each group V and provide the building blocks for the overall estimand. Thus, 2SLS improves upon OLS by identifying the group average effects without weighting issues, but still faces the challenge of appropriately weighting those effects to build the overall estimand.

6 Research Design

In this section, I first present the estimating equations used to make my empirical analysis. I then assess the validity of my instruments in terms of the identifying assumptions discussed in Section 5 and assess the performance of the instruments.

6.1 Estimating Equations

Cross-sectional Estimator: Cumulative Effects

In the first part of my empirical analysis, I use a cross-sectional design to estimate the effect of criminal sentencing on future crime rates of both the defendant and other members of their residence. I measure effects over the first five years after the sentence so that estimates can be interpretted as the cumulative effect of a punishment over this time frame. I define criminal sentences in terms of three different types of punishments: prison sentences, probationary sentences, and fines. The second stage estimating equation is given by

$$\begin{aligned} y_i = {} & \beta_1 Prob_i + \beta_2 Fine_i + \beta_3 PrisDuration_i + \beta_4 (PrisDuration_i)_{<5} \\ & + \beta_5 ProbDuration_i + \beta_6 StayedDuration_i + \boldsymbol{\theta} \boldsymbol{X}_i + \eta_{zip} + \gamma_r + u_i \end{aligned} \tag{3}$$

where criminal cases are indexed by i. The outcome of interest is given by y, which either measures the crimes commited by the defendant or the per-housing unit crime rate at the defendant's address (excluding crimes committed by the defendant). $\boldsymbol{X}$ is a vector of pre-trial defendant characteristics. Crime-by-division-rotation fixed effects are given γ_r and act as the "randomization group" fixed effect. I additionally include zip code fixed effects, η_{zip} to absorb community level variation.

The treatment vector W is comprised of $Prob$ which is an indicator for assignment to probation, $Fine$ which is an indicator for receiving a fine, and the prison, probation, or stayed sentence lengths measured in years. Because outcomes are measured only over the first five years following the sentence, it is expected that the marginal effect of increasing the prison sentence length beyond five years would be zero. Accordingly, I introduce a kink design at the five year mark for prison sentences by including the variable $(PrisDuration)_{<5}$ which takes the value of zero if the prison sentence is longer than five years.

The coefficients β_1 and β_2 give the average effect on the outcome of receiving probation or a fine. β_3 represents the average causal response in the outcome for a marginal increase in the duration of an incarcerated defendant's sentence length while β_4 gives the difference in the average causal response for those sentenced above and below the five year threshold. Similarly, β_5 and β_6 give the average causal response for a marginal increase in the probation and stayed sentence lengths. When treatment effects are heterogeneous, these effects represent weighted averages across defendants of differing severity.

Panel Estimator: Incapacitation and Exposure Effects

In the second part of my analysis, I use a similar research design to Mueller-Smith (2015) in order to disentangle incapacitation effects from exposure effects. To do this, in each year for each defendant I observe whether the defendant is scheduled to be released from their initial court sentence[7] and if so for how long they were exposed to prison or probation. This allows me to construct the following measures for "custody release" and "custody exposure":

$$\begin{aligned} Release_t &= \%\text{ days in custody during year } t \\ Exposure_t &= \text{Number of years in custody through year } t \end{aligned}$$

where $t \in \{0, 1, 2, 3, 4\}$ is the year since sentencing.

$$\begin{aligned} y_{it} = \sum_{k \in \{Prob, Pris\}} &\Big(\beta_{k,1} Release^k_{i,t-1} + \beta_{k,2} Release^k_{it} + \beta_{k,3} Release^k_{it} \times Exposure^k_{it}\Big) \\ &+ \beta_4 Probation + \beta_5 ProbationExposure^k_{it} + \beta W + \theta X_i + \zeta_t + u_{it} \end{aligned} \tag{4}$$

where the outcome y now gives the count of crimes committed by the defendant or the per-housing-unit crime rate committed by others living at the defendant's address in period t. The summation index k is taken over probation and prison so that equivalent terms for probation and prison sentences are included in the estimating equation. The vector W contains all other endogenous variables except for prison and probation sentence lengths from Equation (3). The vector X contains both the covariates and the fixed effects from Equation (3).

[7] Actual releases may differ from scheduled releases when, for example, a defendant receives an early release for good time credit reductions or parole.

Conceptually, the above specification breakdowns sentence lengths into four different groups at the yearly level. First, it differentiates those who are in probation custody and prison custody during period t and the coefficient β_4 represents the takeup effect of probation relative to prison for those still in custody. This effect is moderated by the cumulative exposure to probation, captured by the parameter β_5. Next, the specification differentiates between those who have been released from probation relative to those still on probation. This effect is captured by the released parameters $\beta_{probation,1}$ and $\beta_{probation,2}$, with the interaction parameter $\beta_{probation,3}$ moderating the effect of release by the probation sentence length. Similarly, the effects of prison releases are made as comparisons between those assigned to prison who have been released by year t to those assigned to prison who have not yet been released. The inclusion of lag release measures accounts for dynamics effects in which someone who is arrested for a crime in period $t-1$ is unlikely to commit crime in period t due to incapacitation effects.

Equation (4) also has a close connection to a staggered difference-in-differences estimator (one for probation and one for prison). Here, ζ_t is the time-period fixed effect while the release variables act as the treatment event (with a continuous intensity). The second stage design differs from a more standard staggered difference-in-differences design in that it involves a time-dependent treatment (sentence durations) and there is no treatment group fixed effect term. And while the identification results in Section (5) do not make any assumptions on the form of the second stage, recent research has raised concerns about difference-in-differences research designs in the presence of treatment effect heterogeneity, particularly continuously measured treatments or treatments with heterogeneous effects over time, and so it is worth considering how these results track to the 2SLS setting. Goodman-Bacon (2021) shows that, in the presence of temporal treatment heterogeneity, staggered difference-in-differences estimates a weighted sum of treatment effects in which earlier treatment groups receive a negative weight in later periods. This can be viewed as an omitted treatment bias in which some feature of time drives heterogeneity (for example, timing-based selection into treatment or exposure effects). This is addressed in my analysis by estimating the heterogeneity directly through exposure effects. Difference-in-differences estimators also generally rely on a parallel trends assumption that holds after controlling for time-period and treatment-group fixed effects. Equation (4) does not include any treatment-group fixed effect as this would re-introduce endogeneity based on the defendant's sentence. However, Miller (2023) shows that omitting the treatment-group fixed effect amounts to making an assumption that the treatment is independent of potential outcomes. In this way, the Assumption 1 ensures even the strong parallel trends assumptions for continuous treatments introduced by Callaway et al. (2024). This is discussed in more detail in Appendix C.

6.2 Instrument Construction

As discussed in Section 4, instruments Z are chosen to estimate δ_i in the first stage according to a function that depends on the assigned judge j, committed crime c, and personal characteristics x. I do this with the use of honest forest predictions in the approach of Wager and Athey (2018). I additionally avoid overfitting by training predictions for the potential punishments of i on the sample of defendants that are assigned judge j but do not live in the same zip-code as i (thereby also excluding i from the sample). As a result, the prediction does not depend on the assignment of i or others living in their community.

The results of Wager and Athey (2018) show that honest forest predictions are an unbiased estimator for the conditional mean function of the predicted variable. That is, let X be the set of pre-trial measures used to fit the random forest model and $m(J, X, Zip)$ be the predicted punishment of judge J for an individual with pretrial characteristics X who lives in zipcode zip. Then

$$\mathbb{E}[m(J, X, zip)] = \mathbb{E}[W|J, X, Zip \neq zip].$$

As a comment, the sample selection that omits the defendant's zipcode induces a bias on the random forest predictions and might induce a correlation in the weight distribution and potential outcomes if judges hold strong preferences over individual zip codes. However, the leave-out forest can also be viewed as a cluster JIVE estimator that eliminates many-treatment biases in the second stage that arise when i's instrument is correlated with its own treatment assignment (Frandsen et al., 2024). Thus, this suggests a trade-off between inducing a correlation in the weights over zipcodes and a bias in the second stage outcomes.

The random forest predictions are made over the set of pre-trial characteristics described in Table 1 and additionally dummy indicators for the crime commited by the defendant and a set of time controls to account for temporal dynamics in judge preferences. My instruments are constructed by following the "recentering" approached described in Borusyak and Hull (2023). To do this, I first predict the punishment of defendant i for each possible judge to which they might possibly be assigned to produce a set of predictions $m(J, x_i)$. These predictions are made for judge decisions on probation assignment, probation time, stayed time, prison time, and fine assignment. Then, for each respective regression specification, I construct instruments by first transforming the predicted punishments to obtain the predictions for the endogenous second-stage regression variables, which I denote W^{forest} (for example, in Equation (3) the prediction for $(PrisDuration)_{<5}$ is obtained by transforming the prediction for prison sentence lengths). I then estimate the average predicted regression variables, $\mathbb{E}_J[W^{forest}]$, by averaging the predictions W^{forest}, where the probability of being assigned to judge J is given by the share of defendants assigned to judge J in the division rotation of defendant i. Lastly, I compute the predicted sentencing shock as

$$\widehat{\delta}_i = W^{forest} - \mathbb{E}_J[W^{forest}].$$

By construction, $\widehat{\delta}_i$ is conditionally independent of potential outcomes after conditioning on X and is (unconditionally) mean independent of potential outcomes.

6.3 Testing Identifying Assumptions

The results of Section 5 show that coefficient estimates from Equations (3) and (4) can be interpretted as weighted LATEs only when Assumptions 1, 2, 3, 4, and 5 hold. I assess each of these in turn.

Instrument Exogeneity

Assumption 1 depends on whether the assignment of cases to judges is independent of potential outcomes. This assumption is not directly testable because all potential outcomes are not observed, however indirect balance tests can be conducted by assessing whether assignment is correlated with pre-trial defendant characteristics. Balance tests are commonly conducted by regressing a set of pre-trial characteristics on dummy variables for assignment to each judge. For my setting, the balance test regression is given by

$$X_i = \delta_j + \gamma_r + e_i \tag{5}$$

where δ_j is a fixed effect for judge j, γ_r is a $Crime \times Division \times Cycle$ fixed effect[8], and e_i is a residual error term. The joint F-test for whether $\delta_j = 0$ for all j provides an indirect test for Assumption 1. However, the F-test relies on the assumption that the $\widehat{\delta}_j$ are approximately normally distributed. When the fixed-effect group sizes used to compute $\widehat{\delta}_j$ are small, this assumption generally does not hold and the F-statistic will not be even asymptotically F-distributed (Blanca et al., 2017).

In my setting, assignment groups are formed by the combination of $Judge \times Crime \times Division \times Cycle$. Particularly problematic is the fact that in some cycles are very short (sometimes lasting only a couple months). Because fixed effects for judges j are estimated separately within each rotation cycle, the assignment group sizes do not grow with the sample size and are unequally distributed across randomization blocks. I address this by instead conducting a test based on randomization inference in which I randomly re-assign defendants to a judge within the same randomization block. Under the null hypothesis that the simulated randomization procedure matches the true randomization procedure, the placebo test statistics are drawn from the same distribution as the observed test statistic and inference can then be made by comparing the true statistic against the placebo distribution (Fisher, 1935). This provides a distribution-free test that corrects for the many-group false positive rate. The test procedure is as follows:

1. Estimate Equation (5) and store the resulting F-statistic for joint significance
2. Randomly reassign judges without replacement to a case within the same $Crime \times Division \times Cycle$ block
3. Estimate Equation (5) for the new judge assignment structure created by Step 2
4. Repeat Steps 2 - 3 M times, recording the obtained F-statistic each time
5. Conduct a one-sided test at the α significance level by comparing the F-statistic obtained in Step 1 to the $(1 - \alpha)$ percentile of the M statistics obtained from Steps 2 - 4. The null hypothesis is rejected if the observed statistic is greater than this threshold[9]

Table 2 reports results for these tests against 12 pre-trial defendant characteristics. Each test features a placebo distribution of 100 simulated judge randomizations. In only one case is the null hypothesis rejected (p-value of 0.07). This, however, does not differ dramatically from what would be expected by random chance because over 12 tests it would be expected to reject on average 1.2 tests at the 90% confidence level due to random variation. Thus, I do not find strong evidence against the null that the assignment of defendants to judges is random conditional on the initial $Crime \times Division \times Cycle$ block classification.

[8] The inclusion of γ_r is important because the court assignment procedures discussed above indicate that judge assignment is only random conditional on the initial crime and court division classification within a given court rotation cycle.

[9] The balance test is conducted with a one-sided t-test because selective assignment on pre-trial characteristics should induce greater variation in judge averages than under random assignment. In this case, the F-statistic should grow larger.

Table 2: Placebo Tests for Balance of Randomization

	F-statistic (1)	Placebo 90th Percentile (2)	p-value (3)	N (4)
Units	1.28	1.90	0.63	28,324
Mean unit area	2.84	9.46	0.23	28,324
Mean unit value	1.90	2.09	0.16	28,324
Mean bedrooms	2.84	5.08	0.28	28,324
Age	1.39	5.77	0.75	28,324
Male	1.38	16.56	0.24	28,324
White	6.47	1.92	**0.10**	28,324
Black	7.27	8.32	0.22	28,324
Divorced last year	1.02	1.59	0.78	28,324
Paternity claim last year	1.10	1.73	0.72	28,324
Evicted last year	10.12	9.91	**0.06**	28,324
Committed crime last year	1.37	12.78	0.52	28,324

Notes: Permutation based placebo tests for balance of randomization in judge assignment to criminal defendants. Column (1) presents the F-statistic for joint significance from a regression of defendant pre-trial characteristics on the assigned judge. Column (2) presents the 90th percentile of the placebo distribution of F-statistics obtained from regressions of pre-trial characteristics on judge fixed effects for the re-randomized judicial assignment. Column (3) presents a rank-based p-value obtained from comparing the observed F-statistic of Column (1) to the placebo distribution, with bold values indicating that the observed F-statistic is statisticall different from the placebo distribution at the 10% confidence level. Column (4) presents observation counts for the regressions.

First Stage Relevance

My analysis hinges on the notion that (i) judges differ in their sentencing preferences and (ii) that these preferential differences are predictive of the actual sentences given to defendants.

Table 3 presents summary statistics for actual sentences, random forest predictions of these sentences and the recentered deviation instruments. Approximately 45% of the sample is assigned to probation. Defendants are on average assigned to probation for a length of 0.8 years and with a stayed prison sentenced of 0.73 years and assigned to prison for a length of 1.5 years. Fines are applied in approximately 11% of cases. The random forest predictions very nearly match the first and second moment of the actual distribution, though the RMSE is relatively large for prison sentence lengths, stayed sentences, and fines. The recentered instruments are on average close to zero, consistent with the notion that they are mean-zero variables.

Next, I present the first stage estimates for both the cross-sectional and panel estimators. Appendix Table B3 presents the first stage for the cross-sectional specification given by Equation (3). Each column represents a regression of the indicated sentencing measure on the instruments (rows). The results indicate that in general the first-stage relationship between expected and actual sentencing is strong. Examining the diagonal of the table indicates how a particular dimension of the expected sentence predicts the actual sentence of that same dimension, conditional on the other judge preferences. The diagonal elements are all highly significant, suggesting that the judge's expected punishments are indeed strong predictors for the actual punishment. Sanderson-Windmeijer F-statistics provide a measure for the overall strength of the first stage relationships when considering the degree of common variation across the first stage equations (Sanderson and Windmeijer, 2016). All SW-F statistics range from 151 to over 1,800.

Appendix Table B4 presents the first stage for the panel specification given by Equation (4).

Table 3: Summary Statistics for Average Criminal Punishments

	Punishment (1)	RF Predicted (2)	Recentered IV (3)	RMSE (4)	N
Probation	0.4541 (0.4979)	0.4539 (0.4962)	0.0117 (0.0416)	0.0115	28,324
Probation Time (Years)	0.8038 (1.09)	0.8018 (1.05)	0.0225 (0.1486)	0.1972	28,324
Stayed Sentence (Years)	0.7311 (1.69)	0.7319 (1.56)	0.0215 (0.2676)	0.3992	28,324
Prison Time (Years)	1.52 (5.02)	1.55 (4.60)	0.0729 (1.21)	1.73	28,324
Fine	0.1115 (0.3148)	0.1120 (0.2772)	0.0108 (0.0815)	0.1108	28,324

Notes: Table rows present means and standard deviations below in parentheses for criminal punishments. Column (1) presents statistics for the sample averages. Column (2) presents statistics for random forests predictions of the judge to which the defendant was actually assigned. Column (3) presents statistics for the recentered deviation instruments. Column (4) presents the RMSE of the predictions in column (2) relative to the actual assignments in column (1).

The diagonal is again generally strongly significant and SW-F statistics range from 50 to over 1,100.

Conditional Monotonicity and Exclusion

Bhuller and Sigstad (2023) propose an indirect test to jointly assess the validity of the Assumptions 1, 2, 4, and 5. The null hypothesis is that the assumptions jointly hold and a rejection of the null implies that at least one of the assumptions is not satisfied. The tests then provide an initial sense as to whether the assumptions hold[10]. Because the balance tests conducted above do not provide evidence for imbalance in pre-trial characteristics, I assume that Assumption 1 holds and instead interpret the results as tests of Assumptions 2, 4, and 5. I implement their test as follows:

1. Estimate the first stage for the whole sample
2. Split the sample into two groups on the basis of pre-trial characteristics X
3. Within each group of the sample split, regress actual sentences W onto the fitted values $\widehat{W}$ obtained in Step 1. That is, for each sub-group I regress

$$W_{k,i} = \widetilde{\omega}\widehat{W}_i + \eta_{zip} + \gamma_r + u_i \tag{6}$$

4. For sentencing dimension W_k, conduct a Wald test that the k^{th} coefficent, $\widetilde{\omega}_k$, is non-negative (average conditional monotonicity) and a Wald test for the joint hypothesis that all other coefficients, $\widetilde{\omega}_{-k}$, are not different from zero (conditional exclusion restriction)
5. Repeat Step 4 for each element of W

Appendix Tables B1 and B2 present p-values for these test for the cross-sectional and panel specifications. Each cell represents the p-value from the test within the subgroup indicated by the row and column for the average conditional monotonicity (Panel A) and conditional exclusion

[10] Bhuller and Sigstad (2023) derive the tests within a multiple discrete treatment setting, but their econometric setup closely mirrors that of Section 5.

restriction (Panel B). Bolded p-values indicate a rejection of the null hypothesis at the 10% confidence level. If the null hypothesis is rejected, then the 2SLS estimand may not represent a weighted LATE. For the cross-sectional specification, I reject the null hypothesis for conditional exclusion in 9 cases. Given that I conduct the test across 24 subgroups for 4 treatments (96 total tests), this is not different from what would be expected by random chance. This differs from the panel estimator in which I find strong evidence that the conditional exclusion restriction is violated. This might reflect that the panel estimator struggles to decompose the effect of sentence lengths between exposure, release, and the interaction terms. I find no evidence that the average conditional monotonicity assumption is violated.

As an illustrative exercise, I reconduct the conditional exlcusion restriction tests for Equation (3) when not instrumenting for all elements of W but instead adopt alternative approaches to control for omitted treatments found in the literature. These tests are presented in Appendix E and provide strong evidence that when elements of W are omitted, the exclusion restriction is violated and estimates no longer represent a weighted LATE. Alternative approaches to controlling for omitted treatment biases–by either controlling for the instrument of an omitted treatment or by re-estimating effects in subsamples–do not appear to lead to any improvements in the test.

7 Empirical Results

7.1 Cumulative Effects

Table 4 presents estimates for the effect of criminal sentencing on crimes committed in the subsequent fives years as estimated by Equation (3). Panel A presents effects for counts of crimes committed by the defendant while Panel B presents effects on the crime rate per property unit committed by any other resident of the defendant's address. Each column is a separate regression. For comparison, odd numbered columns provide estimates when sentence durations are omitted while even columns present estimates from my preferred specification when sentence durations are included.

Defendant Responses

First focusing on the extensive margin, the estimates indicate that the marginally probationed defendant commits on average 0.1 more crimes over the first 5 years post-sentencing than if they were incarcerated, which is largely driven by property crimes (increase of 0.07 property crimes), though only the estimated effect on property crimes is different from zero. Studying the effect of sentence durations, I find that both probation and prison sentence lead to statistically significant reductions in crime over the first first years. Specifically, an additional year of probation prevents 0.1 crimes, of which are largely property crimes, while an additional year in prison prevents on average 0.08 crimes, for which I find significant reductions in both violent and non-violent crimes.

Residential Responses

Panel B of Table 4 presents my estimates for effects on crime rates committed by other members of a defendant's residence[11]. In this case, I only find evidence for effects of prison sentence

[11] Rates are computed as the number of crimes committed by someone else living at the same address of the defendant divided by the number of housing unit located at the address.

Table 4: Cross-sectional Effects of Courtroom Sentencing on Crime 5 Years Post-sentencing

	All Crimes		Property Crimes		Violent Crimes		Other Crimes	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Crimes by Defendant								
Extensive Margin								
Probation	0.0511	0.0986	0.0212	0.0738^{*}	0.0132	0.0419	0.0168	-0.0171
	(0.0380)	(0.0691)	(0.0182)	(0.0387)	(0.0213)	(0.0334)	(0.0241)	(0.0443)
Probation Years		-0.1024^{***}		-0.0740^{***}		-0.0235		-0.0049
		(0.0360)		(0.0213)		(0.0174)		(0.0224)
Stayed Years		0.0001		0.0268^{***}		-0.0027		-0.0239^{***}
		(0.0141)		(0.0092)		(0.0066)		(0.0088)
Prison Effects								
$<$ 5 Years		-0.0845^{***}		-0.0238^{**}		-0.0116^{**}		-0.0491^{***}
		(0.0138)		(0.0094)		(0.0057)		(0.0080)
Panel B: Crimes by Others on Property								
Extensive Margin								
Probation	0.0303	0.0608	0.0138	0.0267	0.0078	0.0013	0.0087	0.0328
	(0.0400)	(0.1132)	(0.0145)	(0.0330)	(0.0131)	(0.0359)	(0.0253)	(0.0635)
Probation Years		-0.0787		-0.0284		-0.0070		-0.0433
		(0.0678)		(0.0179)		(0.0217)		(0.0376)
Stayed Years		0.0086		0.0081		-0.0033		0.0038
		(0.0182)		(0.0064)		(0.0064)		(0.0105)
Prison Effects								
$<$ 5 Years		-0.0600^{***}		-0.0156^{**}		-0.0148^{***}		-0.0296^{***}
		(0.0160)		(0.0077)		(0.0051)		(0.0093)
N	28,324	28,324	28,324	28,324	28,324	28,324	28,324	28,324
Kleibergen-Paap rk LM Stat	3,076.59	764.15	3,076.59	764.15	3,076.59	764.15	3,076.59	764.15
Kleibergen-Paap rk F Stat	802.67	206.17	802.67	206.17	802.67	206.17	802.67	206.17
Continuous Effects	No	Yes	No	Yes	No	Yes	No	Yes

Notes: 2SLS estimates for the effect of sentencing on criminal outcomes in the 60 months post-sentencing. Each column is a 2SLS regression of the outcome on sentencing dimensions, instrumented with assigned judge's expected sentence. Panel A presents estimates for crimes commited by the defendant. Panel B presents estimates for crime rates committed by others (excluding the defendant) living at the same residence as the defendant as a rate per property unit on the address. Odd columns exclude continuous dimensions of court sentences from the estimating equation while even columns include continuous sentencing dimensions. Robust standard errors in parentheses. $^{***} = p < 0.01$, $^{**} = p < 0.05$, $^{*} = p < 0.1$.

lengths. Specifically, an additional year in prison reduces the residence crime rate by 0.06 crimes over the first five post-sentencing years. Studying effects across different crime categories, I find that these reductions are driven by both property and violent crimes.

These results might be driven by either deterrent responses to community members sentence or through reduced exposure to an ex-convict. I explore these potential mechanisms in more detail below.

7.2 Mechanisms: Incapacitation and Exposure Effects

The previous analysis provides evidence that both probation and prison reduce crime. Probationed defendants commit more crimes, but this effect can be partially mitigated by extending the length of the probation term. As well, probationed offenders appear most susceptible to commit property crimes. The nature of the punishment also has spillover effects on the local neighborhood of a defendant as longer prison sentences appears to decrease local crime. However, an important policy question is whether these effects purely reflect incapacitation effects associated with higher custody levels or whether incarceration or probation sentences can provide longer lasting behavioral changes that reduce recidivism after the sentence. As discussed in Section 6, these mechanisms can be assessed with the panel estimator given by Equation (4).

Estimates of Equation (4) are presented in Table 5. Panel A again presents results for defendants while Panel B presents results for crimes of other members of the defendant's property as a per household rate. For Panel A, the outcome is the count of crimes commited in year t. For Panel B, the outcome is the number of crimes committed per unit by others living at the same address as the defendant in year t.

Defendant Responses

The results indicate that the marginally probationed defendant commits on average 0.15 more crimes per year than if they were incarcerated. This contrasts with the findings of the cross-sectional estimates that probationed defendants commit on average more crimes in the first five years after the sentence. However, the difference can be explained by the exposure term, which indicates that the increase in crime induced by probation releases is declining as the the offender has been on probation longer. This reveals an underlying dynamic that the cross-sectional estimator cannot identify. Specifically, there is a short-term increase in crime when a defendant is released on probation. However, this causes criminally active defendants to be re-convicted, so that the effect fades over successive years of custody.

Additionally, there is reduction in convictions once the defendant is released from probation. This indicates that part of the increase in arrests associated with probation reflects a policing effect. Specifically, the estimates indicate that a probationed offender is convicted for approximately 0.1 less crimes after being released from probation. This means that while probationed offenders commit more crime than if they were incarcerated, part of the crimes they are convicted for are only identified because of the higher policing levels experienced during the probation sentence (for example, if the offender fails a drug test). Consistent with this, the reductions in crimes driven by releases are largely comprised of the residual other crime category. The magnitude of this release effect is also declining in the level of exposure, indicating again that the effect of probation fades over successive years of custody.

Turning to prison effects, I find that when a defendant is released from prison they commit on average 0.13 more crimes per year, which is driven by both violent and other non-property crimes. The magnitude of these effects nearly match the magnitude of the effects of probation releases, which might indicate that prison incapacitation effects simply act to push incapacitation responses to a later date. There is however no evidence for positive or negative effects for exposure to prison on future crime rates, indicating that increased exposure to prison is neither reformative nor criminogenic for the defendants in my sample.

Residential Responses

I generally find no evidence for residential responses to criminal sentences with the panel estimator. This however is not necessarily inconsistent with the previous findings with the cross-sectional estimator, as Equation (4) only captures a defendant's exposure to different punishments, but it does not reflect the residence's exposure to the defendant.

To address this, I re-estimate the specification given by Equation (4) when additionally including the following endogenous variable:

$$YearsSinceRelease_k = \max(t + 1 - SentenceDuration, 0) \times D_k$$

The measure is defined and included for both prison and probation and is 0 for those note assigned to punishment k or while the defendant is in custody and represents that number of years the defendant has been released from custody for punishment k otherwise.

The results for the effect of prison on residence crime rates when including the "Years Since Release" measures are shown in Table . The results indicate now that while local crime rates decline following the release of an imprisoned defendant, there are increases in crime rates for those with longer setentences. The effects are also increasing the longer that the offender has been returned in the community.

8 Conclusion

In this paper I study how prison sentences and probationary sentences influence the future criminal behavior of both the sentenced defendant and their surrounding community. To identify causal effects, I leverage the random assignment of judges to criminal defendants and apply a 2SLS "judge fixed effect" estimator. In making my analysis, I first provide new econometric results that show that 2SLS estimators for the effect of prison and probation are generally only valid when judge decisions for potential sentence assignments and durations are simultaneously estimated. This framework extends the results of Angrist et al. (2000), Bhuller and Sigstad (2023), Frandsen et al. (2024), and Humphries et al. (2024) by bridging the IV literature on multiple treatments to the continuous treatment setting and more broadly applies to any setting in which a continuous treatment(s) is assigned endogenously, but researchers have access to an instrumental variable that induces exogenous variation in the assignment of both treatment takeup and intensity. These results also formalize the arguments discussed in Mueller-Smith (2015) to better understand the threats to validity in multi-treatment settings under arbitrary first and second stage heterogeneity.

I show that my empirical strategy allows me to overcome omitted treatment biases to a

Table 5: Panel Estimates for the Effects of Sentencing Releases on Crime 5 Years Post-sentencing

	All Crimes (1)	Property Crimes (2)	Violent Crimes (3)	Other Crimes (4)
Panel A: Crimes by Defendant				
Probation Effects				
Probation	0.1494^{***}	0.0223	0.0488^{***}	0.0783^{***}
	(0.0302)	(0.0226)	(0.0149)	(0.0139)
Exposure	-0.0450^{***}	−0.0133	-0.0167^{**}	-0.0149^{**}
	(0.0134)	(0.0095)	(0.0065)	(0.0071)
Release	-0.1048^{***}	−0.0232	−0.0271	-0.0545^{***}
	(0.0326)	(0.0200)	(0.0193)	(0.0168)
Release × Exposure	0.0373^{**}	0.0080	0.0119	0.0174^{**}
	(0.0148)	(0.0080)	(0.0091)	(0.0084)
Prison Effects				
Release	0.1297^{***}	0.0021	0.0506^{**}	0.0769^{***}
	(0.0465)	(0.0367)	(0.0227)	(0.0185)
Release × Exposure	0.0071	0.0264	−0.0173	−0.0020
	(0.0354)	(0.0285)	(0.0173)	(0.0128)
Panel B: Crimes by Others on Property				
Probation Effects				
Probation	0.0440	−0.0171	0.0093	0.0518
	(0.0661)	(0.0236)	(0.0193)	(0.0353)
Exposure	−0.0120	0.0069	0.0003	−0.0191
	(0.0233)	(0.0092)	(0.0071)	(0.0127)
Release	−0.0151	−0.0006	0.0033	−0.0179
	(0.0268)	(0.0176)	(0.0086)	(0.0177)
Release × Exposure	0.0176	0.0009	−0.0009	0.0175^{**}
	(0.0124)	(0.0066)	(0.0046)	(0.0086)
Prison Effects				
Release	−0.0214	−0.0364	−0.0024	0.0175
	(0.0448)	(0.0348)	(0.0102)	(0.0244)
Release × Exposure	0.0285	0.0314	0.0040	−0.0069
	(0.0342)	(0.0273)	(0.0074)	(0.0175)
N	141,715	141,715	141,715	141,715
Kleibergen-Paap rk LM Stat	239.93	239.93	239.93	239.93
Kleibergen-Paap rk F Stat	21.92	21.92	21.92	21.92

Notes: 2SLS estimates corresponding to Equation (4) for the effect of sentencing on criminal outcomes in the year of the observation. Each column is a 2SLS regression of the outcome on sentencing dimensions, instrumented with the assigned judge's expected sentence. Panel A presents estimates for crimes commited by the defendant. Panel B presents estimates for crime rates committed by others (excluding the defendant) living at the same residence as the defendant as a rate per property unit on the address. Standard errors clustered by judge in parentheses. *** = p < 0.01, ** = p < 0.05, * = p < 0.1.

Table 6: Panel Estimates for the Effects of Sentencing Releases on Crime 5 Years Post-sentencing

	All Crimes (1)	Property Crimes (2)	Violent Crimes (3)	Other Crimes (4)
Prison Effects				
Release	-0.6977^{*}	−0.1437	-0.3614^{**}	−0.1925
	(0.4229)	(0.2104)	(0.1684)	(0.2813)
Release × Exposure	0.2243^{*}	0.0614	0.1071^{**}	0.0557
	(0.1259)	(0.0651)	(0.0489)	(0.0832)
Years Since Release	0.2510	0.0455	0.1300^{**}	0.0755
	(0.1528)	(0.0743)	(0.0603)	(0.1008)
N	141,715	141,715	141,715	141,715
Kleibergen-Paap rk LM Stat	26.49	26.49	26.49	26.49
Kleibergen-Paap rk F Stat	2.68	2.68	2.68	2.68

Notes: 2SLS estimates corresponding to Equation (4) for the effect of sentencing on criminal outcomes in the year of the observation. Each column is a 2SLS regression of the outcome on sentencing dimensions, instrumented with the assigned judge's expected sentence. Panel A presents estimates for crimes commited by the defendant. Panel B presents estimates for crime rates committed by others (excluding the defendant) living at the same residence as the defendant as a rate per property unit on the address. Standard errors clustered by judge in parentheses. *** = p < 0.01, ** = p < 0.05, * = p < 0.1.

greater extent than other related studies. My econometric results formalize these arguments and I provide empirical evidence that alternative approaches to controlling for omitted treatment biases are generally less successful than my preferred estimators.

In the first step of my empirical analysis, I find that harsher sentences generally lead to reduced crime. Longer sentences–either probation or prison–reduce future crime rates by the defendants, but the overall reduction in crime from incarceration is larger than probation. In the second part of my analysis, I seek to better understand the mechanisms driving these results.

I find that for prison the effects appear to largely be driven by incapacitation effects as recidivism rates spike immediately within one year of the release, but I find no evidence that increased exposure to prison leads to further reductions (or increases) in crime once the defendant is released. Instead, effects appear to be driven by how rapidly an offender is released from custody, which leads to a relatively persistent gap in crime over the first five years.

For probation, I find that during the probation sentence there is an initial policing effect that appears to increase conviction rates. That is, for those who are assigned to probation and continue to commit crime, the increased levels of supervision act to increase the probability of re-conviction *during* probation. Two-thirds of this effect disappears once the defendant is released from probation.

And while the mechanisms driving prison and probation effects are different in practice, the effects of both are indicative of cyclical patterns of crime in which defendants are caught, punished, released, and caught again. In general, I find little evidence that exposure to either prison or probation leads to desistence, but rather capture different paths through the criminal justice system.

However, I do find consistent evidence that exposure to criminals leads to increases in crime rates at the offender's residence, which is primarily driven by violent crimes. These effects do not appear generally appear to be unique to either probation or prison, as residence crime rates rise when an offender is returned home on probation or returned home after incarceration. This provides evidence that exposure to criminals in the household/local community plays a role in the persistence of criminal behavior within a community.

Overall, these findings highlight policy relevant tradeoffs between incarceration and probation. The direct impact of stringent incarcerative sentences to defendants might be severe through time lost in prison and labor market disruptions, with little long-term reductions in long-run crime from the criminal. However, incarceration does provide greater criminal reductions through the direct incapacitation of the offender and negative spillovers in the community.

References

Aizer, Anna and Joseph J. Doyle, "Juvenile Incarceration, Human Capital, and Future Crime: Evidence from Randomly Assigned Judges," *The Quarterly Journal of Economics*, 2015, *130* (2), 759–804.

Angrist, J. D., G. W. Imbens, and A. B. Krueger, "Jackknife Instrumental Variables Estimation," *Journal of Applied Econometrics*, 1999, *14* (1), 57–67.

Angrist, Joshua D., Kathryn Graddy, and Guido W. Imbens, "The Interpretation of Instrumental Variables Estimators in Simultaneous Equations Models with an Application to the Demand for Fish," *The Review of Economic Studies*, 2000, *67* (3), 499–527.

Arteaga, Carolina, "Parental Incarceration and Children's Educational Attainment," *The Review of Economics and Statistics*, 11 2023, *105* (6), 1394–1410.

Barbarino, Alessandro and Giovanni Mastrobuoni, "The Incapacitation Effect of Incarceration: Evidence from Several Italian Collective Pardons," *American Economic Journal: Economic Policy*, 2014, *6* (1), 1–37.

Bayer, Patrick, Randi Hjalmarsson, and David Pozen, "Building Criminal Capital behind Bars: Peer Effects in Juvenile Corrections," *The Quarterly Journal of Economics*, 2009, *124* (1), 105–147.

Becker, Gary S., "Crime and Punishment: An Economic Approach," *Journal of Political Economics*, 1968, *76* (2), 169–217.

Bhuller, Manudeep and Henrik Sigstad, "2SLS with Multiple Treatments," *Journal of Econometrics (Revise & Resubmit)*, 2023.

_ , Gordon B. Dahl, Katrine V. Loken, and Magne Mogstad, "Intergenerational Effects of Incarceration," *AEA Papers and Proceedings*, May 2018, *108*, 234–40.

_ , _ , Katrine Vellesen Løken, and Magne Mogstad, "Incarceration Spillovers in Criminal and Family Networks," *NBER Working Paper No. w24878*, 2018.

Blanca, María J, Rafael Alarcón, Jaume Arnau, Roser Bono, and Rebecca Bendayan, "Effect of variance ratio on ANOVA robustness: Might 1.5 be the limit?," *Behavior Research Methods*, 2017.

Blandhol, Christine, John Bonney, Magne Mogstad, and Alexander Torgovitsky, "When is TSLS Actually LATE?," *R&R at The Review of Economic Studies*, 2022.

Borusyak, Kirill and Peter Hull, "Nonrandom Exposure to Exogenous Shocks," *Econometrica*, 2023, *91* (6), 2155–2185.

Buonanno, Paulo and Steven Raphael, "Incarceration and Incapacitation: Evidence from the 2006 Italian Collective Pardon," *American Economic Review*, 2013, *103* (6), 2437–2465.

Callaway, Brantly, Andrew Goodman-Bacon, and Pedro H.C. Sant'Anna, "Difference-in-Differences with a Continuous Treatment," 2024.

Carson, E. Ann and Rich Kluckow, "Prisoners in 2022 - Statistical Tables," Technical Report 2023.

Chyn, Eric, Brigham Frandsen, and Emily C. Leslie, "Examiner and Judge Designs in Economics: A Practitioner's Guide," Technical Report, NBER 2024.

Dobbie, Will, Hans Grönqvist, Susan Niknami, Mårten Palme, and Mikael Priks, "The Intergenerational Effects of Parental Incarceration," Working Paper 24186 January 2018.

_ , Jacob Goldin, and Crystal S. Yang, "The Effects of Pretrial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges," *American Economic Review*, February 2018, *108* (2), 201–40.

Drago, Francesco and Roberto Galbiati, "Indirect Effects of a Policy Altering Criminal Behavior: Evidence from the Italian Prison Experiment," *American Economic Journal: Applied Economics*, 2012, *4* (2), 199–218.

_ , _ , and Pietro Vertova, "The Deterrent Effects of Prison: Evidence from a Natural Experiment," *Journal of Political Economy*, 04 2009, *117* (2), 257–280.

Fisher, R.A., *The Design of Experiments*, Oliver & Boyd, 1935.

Frandsen, Brigham, Emily Leslie, and Samuel McIntyre, "Cluster Jackknife Instrumental Variable Estimation," *R&R at Review of Economics and Statistic*, 2024.

_ , Lars Lefgren, and Emily Leslie, "Judging Judge Fixed Effects," *American Economic Review*, January 2023, *113* (1), 253–77.

Goodman-Bacon, Andrew, "Difference-in-differences with variation in treatment timing," *Journal of Econometrics*, 2021, *225* (2), 254–277. Themed Issue: Treatment Effect 1.

Guerino, Paul, Paige M. Harrison, and William J. Sabol, "Prisoners in 2010," Technical Report, Bureau of Justice Statistics 2011.

Heckman, James and Edward Vytlacil, "Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast their Effects in New Environments," in J.J. Heckman and E.E. Leamer, eds., *Handbook of Econometrics*, 1 ed., Vol. 6B, Elsevier, 2007, chapter 71.

Heckman, James J. and Edward Vytlacil, "Structural Equations, Treatment Effects, and Econometric Policy Evaluation1," *Econometrica*, 2005, *73* (3), 669–738.

Humphries, John Eric, Aurelie Ouss, Kamelia Stavreva, Megan T. Stevenson, and Winnie van Dijk, "Conviction, Incarceration, and Recidivism: Understanding the Revolving Door," *Quarterly Journal of Economics (Revise and Resubmit)*, 2024.

Kaeble, Danielle, "Probation and Parole in the United States, 2021," Technical Report 2023.

Kasy, Maximilian, "Instrumental Variables with Unrestricted Heterogeneity and Continuous Treatment," *The Review of Economic Studies*, 06 2014, *81* (4), 1614–1636.

Kessler, Daniel and Steven D. Levitt, "Using Sentence Enhancements to Distinguish Between Deterrence and Incapacitation," *The Journal of Law & Economics*, 1999, *42* (S1), 343–364.

Miller, Douglas L., "An Introductory Guide to Event Study Models," *Journal of Economic Perspectives*, May 2023, *37* (2), 203–30.

Mueller-Smith, Michael, "The Criminal and Labor Market Impacts of Incarceration," *American Economic Review (Revise and Resubmit)*, 2015.

_ **and Kevin T. Schnepel**, "Diversion in the Criminal Justice System," *The Review of Economic Studies*, 07 2020, *88* (2), 883–936.

Norris, Samuel, Matthew Pecenco, and Jeffrey Weaver, "The Effects of Parental and Sibling Incarceration: Evidence from Ohio," *American Economic Review*, September 2021, *111* (9), 2926–63.

Sanderson, Eleanor and Frank Windmeijer, "A weak instrument F-test in linear IV models with multiple endogenous variables," *Journal of Econometrics*, 2016, *190* (2), 212–221. Endogeneity Problems in Econometrics.

Słoczyński, Tymon, "Interpreting OLS Estimands When Treatment Effects Are Heterogeneous: Smaller Groups Get Larger Weights," *The Review of Economics and Statistics*, 05 2022, *104* (3), 501–509.

Wager, Stefan and Susan Athey, "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests," *Journal of the American Statistical Association*, 2018, *113* (523), 1228–1242.

Walmsley, Roy, "World Prison Population List," Technical Report, International Center for Prison Studies 2015.

A Tables

Table B1: Wald Test for the Average Conditional Monotonicity and the Conditional Exclusion Restriction Cross-Sectional Estimator

	Probation		Prison Time		Probation Time		Stayed Time	
	$X = 0$	$X = 1$	$X = 0$	$X = 1$	$X = 0$	$X = 1$	$X = 0$	$X = 1$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Average Conditional Monotonicity								
Multi-unit Residence	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
1{Unit Area > Median}	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
1{Unit Value > Median}	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Family residence	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
1{Age > Median}	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Male	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
White	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Other Race	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Divorced last year	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Paternity claim last year	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Evicted last year	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Committed crime last year	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Panel B: Conditional Exclusion Restriction								
Multi-unit Residence	0.9007	0.9509	0.5934	0.7935	0.7910	0.9170	0.6291	0.7702
1{Unit Area > Median}	0.9496	0.9576	0.8469	0.8878	0.6818	0.8104	0.7390	0.8673
1{Unit Value > Median}	0.5210	0.4577	0.4040	0.2811	0.3313	0.2905	0.5785	0.5582
Family residence	0.5973	0.9531	0.7038	0.9637	0.9316	0.9924	0.8057	0.9363
1{Age > Median}	0.9482	0.9251	0.2105	0.3183	0.6253	0.5445	0.3123	0.1040
Male	**0.0700**	0.7352	**0.0552**	0.9209	**0.0449**	0.7791	0.1645	0.7466
White	0.7296	**0.0444**	0.8619	0.1258	0.5008	**0.0040**	0.4653	**0.0025**
Other Race	0.9986	0.3400	0.9948	0.1019	0.9974	0.2474	0.9964	0.2626
Divorced last year	0.9992	0.6744	0.9970	0.8094	0.9999	0.9811	0.9971	**0.0867**
Paternity claim last year	0.9988	0.7182	0.9988	0.7254	0.9962	**0.0217**	0.9802	**0.0014**
Evicted last year	0.9597	0.6708	0.9889	0.7451	0.9213	0.4909	0.9457	0.3515
Committed crime last year	0.9012	0.1102	0.9896	0.8317	0.9427	0.3242	0.9955	0.7590

Notes: Results from Wald tests for average conditional monotonicity and the conditional exclusion restriction for the cross-sectional estimator given by Equation (3). Rows describe the variable used to determine the sample split, while columns indicate the sub-sample tested on for each sentencing dimension. Each entry indicates a p-value from an F-test that proper weights are satisfied. Bolded entries indicate results that are statistically significant at the 90% significance level. Panel A presents p-values for tests of average conditional monotonicity while Panel B presents p-values for tests of the conditional exclusion restriction.

Table B2: Wald Test for Average Conditional Monotonicity and the Conditional Exclusion Restriction Panel Estimates

	Probation Release		Prison Release		Probation		Fine		Prison Exposure		Probation Exposure		Stayed Sentence	
	$X = 0$	$X = 1$	$X = 0$	$X = 1$	$X = 0$	$X = 1$	$X = 0$	$X = 1$	$X = 0$	$X = 1$	$X = 0$	$X = 1$	$X = 0$	$X = 1$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Panel A: Average Conditional Monotonicity														
Multi-unit Residence	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9989	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
1{Unit Area > Median}	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
1{Unit Value > Median}	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Family residence	0.9999	1.0000	0.9998	1.0000	1.0000	1.0000	0.9918	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
1{Age > Median}	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9993	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Male	0.8636	1.0000	0.9998	1.0000	1.0000	1.0000	0.9837	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
White	1.0000	1.0000	1.0000	0.9994	1.0000	0.9891	1.0000	0.9431	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Other Race	1.0000	0.9420	1.0000	0.9956	1.0000	0.9930	1.0000	0.1263	1.0000	0.7612	1.0000	0.9716	1.0000	0.9870
Divorced last year	1.0000	1.0000	1.0000	0.9420	1.0000	0.3908	1.0000	0.9681	1.0000	0.9996	1.0000	0.9992	1.0000	0.9999
Paternity claim last year	1.0000	1.0000	1.0000	0.9965	1.0000	0.9035	1.0000	0.8576	1.0000	0.9950	1.0000	0.9980	1.0000	0.9997
Evicted last year	1.0000	1.0000	1.0000	0.9870	1.0000	1.0000	1.0000	0.9952	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Committed crime last year	1.0000	1.0000	1.0000	0.9795	1.0000	**0.0916**	1.0000	0.9882	1.0000	1.0000	1.0000	1.0000	1.0000	0.9991
Panel B: Conditional Exclusion Restriction														
Multi-unit Residence	0.7230	0.9389	0.8916	0.9839	0.2642	0.5666	0.7006	0.9359	0.9833	0.9968	0.6553	0.9157	0.7820	0.9402
1{Unit Area > Median}	0.9833	0.9839	0.9837	0.9893	0.8408	0.9231	0.9743	0.9816	0.9389	0.9539	0.9733	0.9725	0.9140	0.9114
1{Unit Value > Median}	0.5729	0.6187	0.6751	0.7528	0.2753	0.3094	0.4632	0.4170	0.1019	**0.0985**	0.4316	0.4626	**0.0090**	**0.0109**
Family residence	0.7329	0.9948	0.8662	0.9983	0.6130	0.9978	0.5441	0.9937	0.9918	1.0000	0.8960	0.9990	0.9714	0.9998
1{Age > Median}	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0010**	**0.0006**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**
Male	**0.0000**	**0.0132**	**0.0000**	**0.0255**	**0.0000**	**0.0000**	**0.0000**	**0.0033**	**0.0000**	**0.0000**	**0.0000**	**0.0272**	**0.0000**	**0.0106**
White	0.2328	**0.0000**	0.3656	**0.0000**	0.9359	**0.0227**	0.8994	**0.0319**	0.7056	**0.0001**	0.6574	**0.0000**	0.6641	**0.0000**
Other Race	1.0000	0.4858	1.0000	0.2017	1.0000	**0.0205**	0.9999	**0.0186**	1.0000	0.2488	1.0000	0.4931	1.0000	**0.0909**
Divorced last year	1.0000	0.4856	1.0000	0.6940	1.0000	0.5855	1.0000	0.3738	1.0000	0.8604	1.0000	0.7272	1.0000	0.9219
Paternity claim last year	1.0000	**0.0795**	1.0000	**0.0731**	1.0000	0.8680	1.0000	0.5540	1.0000	0.8547	1.0000	0.5298	1.0000	0.8403
Evicted last year	0.9210	**0.0000**	0.9833	**0.0031**	0.9874	0.1125	0.9854	**0.0681**	0.9933	**0.0172**	0.9073	**0.0000**	0.9881	**0.0177**
Committed crime last year	0.6992	**0.0000**	0.2908	**0.0000**	0.6606	**0.0000**	0.6987	**0.0025**	0.4556	**0.0000**	0.6018	**0.0000**	0.6225	**0.0000**

Notes: Results from Wald tests for average conditional monotonicity and the conditional exclusion restriction for the SDIVD estimator given by Equation (4). Rows describe the variable used to determine the sample split, while columns indicate the sub-sample tested on for each sentencing dimension. Each entry indicates a p-value from an F-test that proper weights are satisfied. Bolded entries indicate results that are statistically significant. Panel A presents p-values for tests of average conditional monotonicity while Panel B presents p-values for tests of the conditional exclusion restriction.

Table B3: First Stage Estimates
Cross-sectional Estimates

	Probation (1)	Probation Time (2)	Stayed Sentence (3)	Prison Time$_{<5}$ (4)	Fine (5)
$\Delta\{\text{Probation}\|X\}$	5.71^{***}	6.55^{***}	3.02^{***}	-3.43^{***}	-0.4199^{***}
	(0.1161)	(0.2525)	(0.5627)	(0.1857)	(0.0502)
$\Delta\{\text{Probation Time}\|X\}$	-0.0802^{***}	1.75^{***}	1.22^{***}	-0.0939^{***}	0.0088
	(0.0264)	(0.1004)	(0.2727)	(0.0286)	(0.0086)
$\Delta\{\text{Stayed Sentence}\|X\}$	0.0359^{***}	0.1880^{***}	1.89^{***}	-0.0177	-0.0018
	(0.0120)	(0.0577)	(0.2156)	(0.0141)	(0.0037)
$\Delta\{\text{Prison Time}_{<5}\|X\}$	0.0394^{**}	0.0102	-0.0753^{**}	0.6463^{***}	-0.0614^{***}
	(0.0154)	(0.0247)	(0.0332)	(0.0715)	(0.0128)
$\Delta\{\text{Fine}\|X\}$	-0.4737^{***}	-0.7785^{***}	-0.5116^{***}	0.7690^{***}	2.07^{***}
	(0.0333)	(0.0692)	(0.0855)	(0.1090)	(0.0596)
N	28,324	28,324	28,324	28,324	28,324
SW-F	1,853	579	151	873	1,830
Crime-Division-Rotation FEs	Yes	Yes	Yes	Yes	Yes
Pre-Trial Controls	Yes	Yes	Yes	Yes	Yes

Notes: First-stage estimates for Equation (3) for the relationship between the expected sentencing instruments and endogenous variables of actual sentencing. Each column is a regression of the endogenous variable on all instruments. SW-F statistics represent Sanderson-Windmeijer F-statistics for tests of weak instruments with multiple endogenous variables. Standard errors clustered by judge in parentheses. *** = $p < 0.01$, ** = $p < 0.05$, * = $p < 0.1$.

Table B4: First Stage Estimates
Panel Estimates

	$\text{Probation Release}_{t-1}$	$\text{Prison Release}_{t-1}$	Probation	Fine	Prison Custody	Probation Custody	Prison Exposure	Probation Exposure	Stayed Sentence
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Extensive Margin									
$\mathbb{E}\{\text{Probation Release}_{t-1}\|J\}$	-0.0439^{***}	0.8304^{***}	0.0060	0.0262^{*}	-0.0505^{***}	0.1261^{***}	-0.4789^{***}	0.7047^{***}	-0.0762^{*}
	(0.0160)	(0.0236)	(0.0133)	(0.0143)	(0.0171)	(0.0314)	(0.0322)	(0.0839)	(0.0412)
$\mathbb{E}\{\text{Prison Release}_{t-1}\|J\}$	0.8091^{***}	-0.0219	-0.0950^{***}	0.0814^{***}	-0.1641^{***}	0.1020^{***}	0.0891^{*}	-0.2568^{***}	-0.1400^{**}
	(0.0495)	(0.0262)	(0.0200)	(0.0294)	(0.0328)	(0.0315)	(0.0473)	(0.0449)	(0.0704)
$\mathbb{E}\{\text{Probation}\|J\}$	0.1070^{***}	-0.1168^{***}	0.5985^{***}	0.0887^{***}	-0.1436^{***}	-0.1715^{***}	0.5530^{***}	-0.0719	0.3510^{***}
	(0.0184)	(0.0206)	(0.0790)	(0.0274)	(0.0399)	(0.0357)	(0.0660)	(0.0672)	(0.1182)
$\mathbb{E}\{\text{Fine}\|J\}$	-0.0022	0.0806^{**}	0.1203^{**}	0.8044^{***}	-0.0006	0.0323^{*}	0.0102	0.0488	-0.0689
	(0.0311)	(0.0338)	(0.0583)	(0.0440)	(0.0170)	(0.0184)	(0.0245)	(0.0326)	(0.0688)
Prison Durations									
$\mathbb{E}\{\text{Custody (Prison)}_t\|J\}$	0.0396^{***}	-0.0208^{**}	-0.0144	0.0247^{***}	0.0424^{***}	0.0173^{*}	0.0172^{***}	-0.0276^{*}	0.0521^{***}
	(0.0071)	(0.0087)	(0.0100)	(0.0090)	(0.0075)	(0.0099)	(0.0055)	(0.0161)	(0.0162)
$\mathbb{E}\{\text{Prison Time}_t\|J\}$	0.0031	-0.0019	-0.0038	0.0055^{***}	-0.0070^{*}	0.0938^{***}	-0.0060^{*}	0.0583^{***}	-0.0093
	(0.0022)	(0.0025)	(0.0029)	(0.0021)	(0.0037)	(0.0113)	(0.0031)	(0.0116)	(0.0155)
Probation Durations									
$\mathbb{E}\{\text{Custody (Probation)}_t\|J\}$	-0.0708^{***}	-0.0226^{***}	-0.0157	0.0297^{***}	0.0221^{**}	-0.0829^{***}	0.1878^{***}	-0.2875^{***}	-0.0922^{*}
	(0.0184)	(0.0077)	(0.0114)	(0.0114)	(0.0110)	(0.0121)	(0.0186)	(0.0370)	(0.0472)
$\mathbb{E}\{\text{Probation Time}_t\|J\}$	-0.0307^{***}	0.0012	-0.0014	0.0050^{*}	-0.0294^{***}	0.0802^{***}	-0.0236^{***}	0.2236^{***}	-0.0055
	(0.0059)	(0.0035)	(0.0036)	(0.0030)	(0.0045)	(0.0082)	(0.0044)	(0.0181)	(0.0148)
$\mathbb{E}\{\text{Stayed Time}\|J\}$	0.0077^{*}	0.0110^{***}	-0.0444^{***}	-0.0001	0.0093	-0.0026	-0.0487^{***}	0.0205^{*}	0.1618^{***}
	(0.0044)	(0.0041)	(0.0120)	(0.0038)	(0.0082)	(0.0076)	(0.0074)	(0.0117)	(0.0540)
N	141,435	141,435	141,435	141,435	141,435	141,435	141,435	141,435	141,435
SW-F	243	537	641	739	293	85.09	1,140	160	50.60
Crime-Division-Rotation FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Pre-Trial Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: First-stage estimates for Equation (3) for the relationship between the expected sentencing instruments and endogenous variables of actual sentencing. Each column is a regression of the endogenous variable on all instruments. SW-F statistics represent Sanderson-Windmeijer F-statistics for tests of weak instruments with multiple endogenous variables. Standard errors clustered by judge in parentheses. $^{***} = p < 0.01$, $^{**} = p < 0.05$, $^{*} = p < 0.1$.

B Econometric Results

Lemma 5. *Under Assumption 3, β^{2SLS} is given by*

$$\beta^{2SLS} = \sum_{k=1}^{K} Cov(\widehat{W}, WD_k)^{-1} \sum_{k=1}^{K} Cov(\widehat{W}, g^k(T_k, U)D_k).$$

or equivalently by

$$\beta^{2SLS} = \sum_{k=1}^{K} Var(\widehat{W})^{-1} \sum_{k=1}^{K} Cov(\widehat{W}, g^k(T_k, U)D_k).$$

where $\widehat{W}$ is given by

$$\widehat{W} := Var(Z)^{-1}Cov(Z, W)Z$$

and $g^k(T_k, U)$ is the potential outcome when receiving treatment k with a dosage of T_k.

Proof. Under Assumption 3, $Cov(Z, W)^{-1}$ exists. It follows that

$$\begin{aligned}
Var(\widehat{W})^{-1}Cov(\widehat{W}, Y) &= Cov(\widehat{W}, \widehat{W})^{-1}Cov(\widehat{W}, Y) \\
&= Cov(\widehat{W}, W - (W - \widehat{W}))^{-1}Cov(\widehat{W}, Y) \\
&= [Cov(\widehat{W}, W) - Cov(\widehat{W}, W - \widehat{W})]^{-1}Cov(\widehat{W}, Y) \\
&= Cov(\widehat{W}, W)^{-1}Cov(\widehat{W}, Y) \\
&= Cov\Big(Var(Z)^{-1}Cov(Z, W)Z, W\Big)^{-1} Cov\Big(Var(Z)^{-1}Cov(Z, W)Z, Y\Big) \\
&= Cov\Big(Z, W\Big)^{-1} (Var(Z)^{-1}Cov(Z, W))^{-1}Var(Z)^{-1}Cov(Z, W)Cov\Big(Z, Y\Big) \\
&= Cov(Z, W)^{-1}Cov(Z, Y) \\
&= \beta^{2SLS}.
\end{aligned}$$

The third equality holds from the distributive property of covariances. The fourth equality holds from the fact that $W - \widehat{W}$ is the residual of projection of W on Z and hence is uncorrelated with the projection itself, given by $\widehat{W}$. The fifth equality follows from replacing $\widehat{W}$ with its definition. The sixth equality holds as a property of the linearity property of covariance matrices and as a property of the inverse of products of matrices.

Now focus on the term $Cov(\widehat{W}, Y)$. Then note that this may equivalently be written as

$$\begin{aligned}
Cov(\widehat{W}, Y) &= Cov\Big(\widehat{W}, \sum_{k=1}^{K} g^k(T_k, U)D_k\Big) \\
&= \sum_{k=1}^{K} Cov\big(\widehat{W}, g^k(T_k, U)D_k\big)
\end{aligned}$$

Similarly, we have that

$$Cov(\widehat{W}, W) = \sum_{k=1}^{K} Cov\big(\widehat{W}, WD_k\big).$$

□

Lemma 6. *Under Assumptions 1 and 2, the covariance of $Cov(\widehat{W}, Y|D_k = 1)$ is given by*

$$Cov(\widehat{W}, Y|k, v) = \int_0^\infty \beta_{k,v}^{ACRT}(t)\ Var\big(\mathbb{1}\{T_k > t\}|k, v\big)\Big(\mathbb{E}[\widehat{W}|k, v, T_k > t] - \mathbb{E}[\widehat{W}|k, v, T_k \leq t]\Big)dt.$$

Proof. Consider the conditional covariance $Cov(\widehat{W}, Y|D_k = 1, v)$. Expressing Y under treatment k in terms of its potential outcome function gives

$$Cov\Big(\widehat{W}, Y|D_k = 1, v\Big) = Cov\Big(\widehat{W}, g^k(T_k, U)|k, v\Big).$$

Now we may write $g^k(T_k, U)$ as

$$g^k(T_k, U) = g^k(0, U) + \int_0^{T_k} \frac{\partial}{\partial t} g^k(t, U)dt = g^k(0, U) + \int_0^\infty \frac{\partial}{\partial t} g^k(t, U)\mathbb{1}\{T_k > t\}dt.$$

Thus, the conditional covariance $Cov(\widehat{W}, Y|k, v)$ becomes

$$\begin{aligned}
Cov(\widehat{W}, Y, |k, v) &= Cov\Big(\widehat{W},\ g^k(0, U) + \int_0^\infty \frac{\partial}{\partial t} g^k(t, U)\mathbb{1}\{T_k > t\}dt\ |k, v\Big) \\
&= Cov\Big(\widehat{W},\ g^k(0, U)\ |k, v\Big) + Cov\Big(\widehat{W}, \int_0^\infty \frac{\partial}{\partial t} g^k(t, U)\mathbb{1}\{T_k > t\}dt\ |k, v\Big) \\
&= Cov\Big(\widehat{W}, \int_0^\infty \frac{\partial}{\partial t} g^k(t, U)\mathbb{1}\{T_k > t\}dt\ |k, v\Big) \\
&= \mathbb{E}\Big[(\widehat{W} - \mathbb{E}[\widehat{W}|k, v]) \int_0^\infty \frac{\partial}{\partial t} g^k(t, U)\mathbb{1}\{T_k > t\}dt\ \Big|k, v\Big] \\
&= \int_0^\infty \mathbb{E}\Big[\frac{\partial}{\partial t} g^k(t, U)\mathbb{1}\{T_k > t\}(\widehat{W} - \mathbb{E}[\widehat{W}|k, v])\ \Big|k, v\Big]dt \\
&= \int_0^\infty \mathbb{E}\Big[\frac{\partial}{\partial t} g^k(t, U)\Big|k, v\Big]\mathbb{E}\Big[\mathbb{1}\{h^k(Z, V) > t\}(\widehat{W} - \mathbb{E}[\widehat{W}|k, v])\ \Big|k, v\Big]dt \\
&= \int_0^\infty \beta_{k,v}^{ACRT}(t)\ \mathbb{E}\Big[\mathbb{1}\{h^k(Z, V) > t\}(\widehat{W} - \mathbb{E}[\widehat{W}|k, v])\ \Big|k, v\Big]dt \\
&= \int_0^\infty \beta_{k,v}^{ACRT}(t)\ \mathbb{E}[\mathbb{1}\{T_k > t\}(\widehat{W} - \mathbb{E}[\widehat{W}])|k, v]dt.
\end{aligned}$$

The third equality holds from Assumptions 1 and 2. The fifth equality holds from Fubini's Theorem. The sixth equality follows from Assumptions 1 and 2. Specifically, after conditioning on v, $\mathbb{1}\{h^k(Z, V) > t\}(\widehat{W} - \mathbb{E}[\widehat{W}|k, v])$ is simply a function of Z. Assumptions 1 and 2 then imply that $g^k(t, U)$ is independent of Z (and any function of Z) and hence the expectation can be separated. The seventh equality applies the definition of $\beta_{k,h}^{ACRT}(t)$.

Now consider the expectation term given by $\mathbb{E}\Big[\mathbb{1}\{T_k > t\}(\widehat{W} - \mathbb{E}[\widehat{W}])\ \Big|k, v\Big]$. This can be rewritten as

$$\mathbb{E}\Big[\mathbb{1}\{T_k > t\}(\widehat{W} - \mathbb{E}[\widehat{W}|k, v])\ \Big|k, v\Big] = \mathbb{P}\big(T_k > t|k, v\big)\Big(\mathbb{E}[\widehat{W}|k, v, T_k > t] - \mathbb{E}[\widehat{W}|k, v]\Big).$$

Thus, we may express the conditional covariance $Cov(\widehat{W}, Y, |k, v)$

$$\int_0^\infty \beta_{k,v}^{ACRT}(t)\mathbb{P}(T_k > t|k, v)\Big(\mathbb{E}[\widehat{W}|k, v, T_k > t] - \mathbb{E}[\widehat{W}|k, v]\Big)dt \tag{7}$$

The term $\mathbb{E}[\widehat{W}|k,v]$ may be rewritten as

$$\mathbb{E}[\widehat{W}|k,v] = \mathbb{P}\big(T_k > t|k,v\big)\mathbb{E}[\widehat{W}|k,v,T_k > t] + \big(1 - \mathbb{P}(T_k > t|k,v)\big)\mathbb{E}[\widehat{W}|k,v,T_k \leq t]$$

If follows then that

$$\begin{aligned}&\mathbb{E}\Big[\mathbb{1}\{T_k > t\}(\widehat{W} - \mathbb{E}[\widehat{W}|k,v])\ \big|k,v\Big]\\ &\quad = \mathbb{P}\big(T_k > t|k,v\big)\Big(1 - \mathbb{P}\big(T_k > t|k,v\big)\Big)\Big(\mathbb{E}[\widehat{W}|k,v,T_k > t] - \mathbb{E}[\widehat{W}|k,v,T_k \leq t]\Big)\\ &\quad = Var\big(\mathbb{1}\{T_k > t\}|k,v\big)\Big(\mathbb{E}[\widehat{W}|k,v,T_k > t] - \mathbb{E}[\widehat{W}|k,v,T_k \leq t]\Big).\end{aligned}$$

This gives the result that Formula (7) is equivalently given as

$$Cov(\widehat{W}, Y|k,v) = \int_0^\infty \beta^{ACRT}_{k,v}(t)\ Var\big(\mathbb{1}\{T_k > t\}|k,v\big)\Big(\mathbb{E}[\widehat{W}|k,v,T_k > t] - \mathbb{E}[\widehat{W}|k,v,T_k < t]\Big)dt.$$

□

Lemma 7. *The summed covariances of the conditional expectations of $\widehat{W}$ and potential outcomes, conditional on D_k and v are given by*

$$\sum_{k=1}^K Cov\Big(\mathbb{E}[\widehat{W}|D_k,v], \mathbb{E}[g^k(T_k,U)D_k|D_k,v]\Big) = \mathbb{E}\Big[Cov\Big(\mathbb{E}[\widehat{W}|D], \mathbb{E}[h(Z)|D]\ |v\Big)\ \mathbb{E}[\beta^{level}|v]\Big]$$

Proof. First consider the term $Cov\Big(\mathbb{E}[\widehat{W}|D_k,v], \mathbb{E}[g^k(T_k,U)D_k|D_k,v]\Big)$. This can be equivalently expressed as

$$\begin{aligned}Cov\Big(\mathbb{E}[\widehat{W}|D_k,v], \mathbb{E}[g^k(T_k,U)D_k|D_k,v]\Big) &= \mathbb{E}\Big[(\mathbb{E}[\widehat{W}|D_k,v] - \mathbb{E}[\widehat{W}])\mathbb{E}[g^k(T_k,U)D_k|D_k,v]\Big]\\ &= \sum_v \mathbb{P}(D_k = 1, v)(\mathbb{E}[\widehat{W}|k,v] - \mathbb{E}[\widehat{W}])\ \mathbb{E}[g^k(T_k,U)|k,v].\end{aligned}$$

Now consider the term $\mathbb{E}[g^k(T_k,U)|k,v]$. This term represents the expected potential outcomes for response type v who receive treatment k. We may express this generally as a deviation from the mean outcome for type v:

$$\begin{aligned}\mathbb{E}[g(W,U)|k,v] &= (\mathbb{E}[g(W,U)|k,v] - \mathbb{E}[g(W,U)|v]) + \mathbb{E}[g(W,U)|v]\\ &= (1 - \mathbb{P}(k|v))\Big(\mathbb{E}[g(W,U)|k,v] - \mathbb{E}[g(W,U)|D_k \neq 1, v]\Big) + \mathbb{E}[g(W,U)|v]\\ &= (1 - \mathbb{P}(k|v))\mathbb{E}[T_k|k,v]\beta^{takeup}_{k,v} + \mathbb{E}[g(W,U)|v].\end{aligned}$$

Returning to the covariance term, we then have

$$\begin{aligned}&Cov\Big(\mathbb{E}[\widehat{W}|D_k,v], \mathbb{E}[g^k(T_k,U)D_k|D_k,v]\Big)\\ &\quad = \sum_v \mathbb{P}(k,v)(\mathbb{E}[\widehat{W}|k,v] - \mathbb{E}[\widehat{W}])\ \Big((1 - \mathbb{P}(k|v))\mathbb{E}[T_k|k,v]\beta^{takeup}_{k,v} + \mathbb{E}[g(W,U)|v]\Big)\end{aligned}$$

This may be expressed as the sum of two summations. The first summation is given by

$$\begin{aligned}
&\sum_v (\mathbb{E}[\widehat{W}|k,v] - \mathbb{E}[\widehat{W}])\mathbb{P}(k,v)\ (1-\mathbb{P}(k|v))\mathbb{E}[T_k|k,v]\beta^{takeup}_{k,v} \\
&\quad = \sum_v (\mathbb{E}[\widehat{W}|k,v] - \mathbb{E}[\widehat{W}])\mathbb{P}(v)Var(D_k=1|v)\ \mathbb{E}[T_k|k,v]\beta^{takeup}_{k,v} \\
&\quad = \sum_v \mathbb{P}(v)\kappa_{k,v}\beta^{takeup}_{k,v} \\
&\quad = \mathbb{E}_v\Big[\kappa_{k,v}\beta^{takeup}_{k,v}\Big].
\end{aligned}$$

where $\mathbb{E}_v$ indicates that the expectation is taken over v and

$$\kappa_{k,v} := (\mathbb{E}[\widehat{W}|k,v] - \mathbb{E}[\widehat{W}])Var(D_k=1|v)\mathbb{E}[T_k|k,v].$$

The second summation is given by

$$\sum_v \mathbb{P}(k,v)(\mathbb{E}[\widehat{W}|k,v] - \mathbb{E}[\widehat{W}])\mathbb{E}[g(W,U)|v].$$

The covariance then is given by

$$\begin{aligned}
&Cov\Big(\mathbb{E}[\widehat{W}|D_k,v], \mathbb{E}[g^k(T_k,U)D_k|D_k,v]\Big) \\
&\quad = \mathbb{E}_v\Big[\kappa_{k,v}\beta^{takeup}_{k,v}\Big] + \sum_v \mathbb{P}(k,v)(\mathbb{E}[\widehat{W}|k,v] - \mathbb{E}[\widehat{W}])\mathbb{E}[g(W,U)|v]
\end{aligned}$$

Taking the summation over all k treatments gives

$$\begin{aligned}
&\sum_{k=1}^{K} Cov\Big(\mathbb{E}[\widehat{W}|D_k,v], \mathbb{E}[g^k(T_k,U)D_k|D_k,v]\Big) \\
&\quad = \sum_{k=1}^{K}\mathbb{E}\Big[\kappa_{k,v}\beta^{takeup}_{k,v}\Big] + \sum_{k=1}^{K}\sum_v \mathbb{P}(k,v)(\mathbb{E}[\widehat{W}|k,v] - \mathbb{E}[\widehat{W}])\mathbb{E}[g(W,U)|v] \\
&\quad = \sum_{k=1}^{K}\mathbb{E}\Big[\kappa_{k,v}\beta^{takeup}_{k,v}\Big] + Cov\Big(\mathbb{E}[\widehat{W}|K,V], \mathbb{E}[g(W,U)|V]\Big).
\end{aligned}$$

The last step is to show that under Assumption 1,$Cov\Big(\mathbb{E}[\widehat{W}|K,V], \mathbb{E}[g(W,U)|V]\Big) = 0$.

$$\begin{aligned}
Cov\Big(\mathbb{E}[\widehat{W}|K,V], \mathbb{E}[g(W,U)|V]\Big) &= \mathbb{E}\Big[\mathbb{E}[\widehat{W}|K,V]\ \mathbb{E}[g(W,U)|V]\Big] - \mathbb{E}[\widehat{W}]\mathbb{E}[g(W,U)] \\
&= \mathbb{E}\Big[\mathbb{E}\Big[\mathbb{E}[\widehat{W}|K,V]\ \mathbb{E}[g(W,U)|V]\ \Big|V\Big]\Big] - \mathbb{E}[\widehat{W}]\mathbb{E}[g(W,U)] \\
&= \mathbb{E}\Big[\mathbb{E}\Big[\mathbb{E}[\widehat{W}|K,V]\ \Big|V\Big]\ \mathbb{E}[g(W,U)|V]\Big] - \mathbb{E}[\widehat{W}]\mathbb{E}[g(W,U)] \\
&= \mathbb{E}\Big[\mathbb{E}[\widehat{W}|V]\ \mathbb{E}[g(W,U)|V]\Big] - \mathbb{E}[\widehat{W}]\mathbb{E}[g(W,U)] \\
&= \mathbb{E}\Big[\mathbb{E}[\widehat{W}]\ \mathbb{E}[g(W,U)|V]\Big] - \mathbb{E}[\widehat{W}]\mathbb{E}[g(W,U)] \\
&= \mathbb{E}[\widehat{W}]\mathbb{E}[g(W,U)] - \mathbb{E}[\widehat{W}]\mathbb{E}[g(W,U)] \\
&= 0
\end{aligned}$$

where the second, third, fourth, and sixth equalities apply the law of iterated expectations and the fifth equality applies Assumption 1. □

Theorem 1. *Under Assumption 1, 2, and 3, the 2SLS estimand is given by*

$$\beta^{2SLS} = \boldsymbol{\omega}^{-1}\sum_{k=1}^{K}\int_0^{\infty}\mathbb{E}[\tau_{k,v}(t)\beta_{k,v}^{ACRT}(t)]dt + \boldsymbol{\omega}^{-1}\sum_{k=1}^{K}\mathbb{E}\Big[\kappa_{k,v}\beta_{k,v}^{takeup}\Big]$$

where

$$\begin{aligned}\boldsymbol{\omega} &\coloneqq \sum_{k=1}^{K}\int_0^{\infty}\mathbb{E}[\lambda_{k,v}(t)]dt + \sum_{k=1}^{K}\mathbb{E}[\kappa_{k,v}]\\ \tau_{k,v}(t) &\coloneqq Var(\mathbb{1}\{T_k > t\}|v)\Big(\mathbb{E}[\widehat{W}|k,v,T_k > t] - \mathbb{E}[\widehat{W}|k,v,T_k \leq t]\Big)\\ \kappa_{k,v} &\coloneqq Var(D_k|v)\Big(\mathbb{E}[\widehat{W}|k,v] - \mathbb{E}[\widehat{W}]\Big)\mathbb{E}[T_k|k,v]\end{aligned}$$

Proof. From Lemma 5, the 2SLS estimand is given by

$$\begin{aligned}\beta^{2SLS} &= Cov(\widehat{W},W)^{-1}Cov(\widehat{W},Y)\\ &= \Big(\sum_{k=1}^{K}Cov(\widehat{W},h^k(Z,V)D_k)\Big)^{-1}\sum_{k=1}^{K}Cov(\widehat{W},g^k(T_k,U)D_k)\end{aligned}$$

Applying the Law of Total Covariance to each of the covariance terms gives

$$\begin{aligned}Cov(\widehat{W},W) &= \sum_{k=1}^{K}\mathbb{E}[Cov(\widehat{W},h^k(Z,V)D_k|D_k,v)] + \sum_{k=1}^{K}Cov\Big(\mathbb{E}[\widehat{W}|D_k,v],\mathbb{E}[h^k(Z,V)D_k|D_k,v]\Big)\\ Cov(\widehat{W},Y) &= \sum_{k=1}^{K}\mathbb{E}[Cov(\widehat{W},g^k(T_k,U)D_k|D_k,v)] + \sum_{k=1}^{K}Cov\Big(\mathbb{E}[\widehat{W}|D_k,v],\mathbb{E}[g^k(T_k,U)D_k|D_k,v]\Big)\end{aligned}$$

Consider that when $D_k = 0$, the terms $h^k(Z,V)D_k$ and $g^k(T_k,U)D_k$ are also always zero. Hence the conditional covariances are also zero for any v when $D_k = 0$:

$$\begin{aligned}Cov(\widehat{W},h^k(Z,V)D_k|D_k = 0,v) = 0\\ Cov(\widehat{W},g^k(T_k,U)D_k|D_k = 0,v) = 0\end{aligned}$$

It follows then that we may write the unconditional covariances above as

$$\begin{aligned}Cov(\widehat{W},W) &= \sum_{k=1}^{K}\sum_{v}\mathbb{P}(k,v)Cov\Big(\widehat{W},h^k(Z,V)D_k|k,v\Big)\\ &\quad+\sum_{k=1}^{K}Cov\Big(\mathbb{E}[\widehat{W}|D_k,v],\mathbb{E}[h^k(Z,V)D_k|D_k,v]\Big)\\ Cov(\widehat{W},Y) &= \sum_{k=1}^{K}\sum_{v}\mathbb{P}(k,v)Cov\Big(\widehat{W},g^k(T_k,U)D_k|k,v\Big)\\ &\quad+\sum_{k=1}^{K}Cov\Big(\mathbb{E}[\widehat{W}|D_k,v],\mathbb{E}[g^k(T_k,U)D_k|D_k,v]\Big)\end{aligned}$$

Now consider $Cov(\widehat{W}, Y)$. By Lemma 6 and Lemma 7, this is equivalently given by

$$
\begin{aligned}
Cov(\widehat{W}, Y) &= \sum_{k=1}^{K}\sum_{v}\mathbb{P}(k,v)\int_0^\infty \beta_{k,v}^{ACRT}(t)\ Var\big(\mathbb{1}\{T_k > t\}|k,v\big)\Big(\mathbb{E}[\widehat{W}|k,v,T_k > t] - \mathbb{E}[\widehat{W}|k,v,T_k < t]\Big)dt \\
&\quad + \sum_{k=1}^{K}\mathbb{E}\Big[\kappa_{k,v}\beta_{k,v}^{takeup}\Big] \\
&= \sum_{k=1}^{K}\sum_{v}\mathbb{P}(v)\int_0^\infty \tau_{k,v}(t)\beta_{k,v}^{ACRT}(t)dt + \sum_{k=1}^{K}\mathbb{E}\Big[\kappa_{k,v}\beta_{k,v}^{takeup}\Big] \\
&= \sum_{k=1}^{K}\int_0^\infty \mathbb{E}\Big[\tau_{k,v}(t)\beta_{k,v}^{ACRT}(t)\Big]dt + \sum_{k=1}^{K}\mathbb{E}\Big[\kappa_{k,v}\beta_{k,v}^{takeup}\Big].
\end{aligned}
$$

where $\tau_{k,v}(t) \coloneqq Var\big(\mathbb{1}\{T_k > t\}|v\big)\Big(\mathbb{E}[\widehat{W}|k,v,T_k > t] - \mathbb{E}[P|k,v,T_k \leq t]\Big)$.

Now consider $Cov(\widehat{W}, W)$. By a similar argument to Lemma 6, we can define $T_k = h^k(Z,V)$ as

$$T_k = \int_0^\infty \mathbb{1}\{t < T_k\}dt.$$

It follows that

$$\sum_{k=1}^{K}\sum_{v}\mathbb{P}(k,v)Cov(\widehat{W}, h^k(Z,V)D_k|k,v) = \sum_{k=1}^{K}\int_0^\infty \mathbb{E}[\tau_{k,v}(t)]dt.$$

Similarly, by setting $g^k(W,U) = h^k(Z,V)$ in Lemma 7, it follows that

$$\sum_{k=1}^{K} Cov\Big(\mathbb{E}[\widehat{W}|D_k,v], \mathbb{E}[h^k(Z,V)D_k|D_k,v]\Big) = \sum_{k=1}^{K}\mathbb{E}[\kappa_{k,v}].$$

We have then that

$$Cov(\widehat{W}, W) = \sum_{k=1}^{K}\int_0^\infty \mathbb{E}[\tau_{k,v}(t)]dt + \sum_{k=1}^{K}\mathbb{E}[\kappa_{k,v}].$$

□

Theorem 2. *Under Assumptions 1, 2, and 3, the 2SLS parameter* β_k^{2SLS} *is given by*

$$\beta_k^{2SLS} = \omega^{-1}\sum_{l=1}^{K}\int_0^\infty \mathbb{E}\Big[\beta_{l,v}^{ACRT}(t)\tau_{l,v}(t)\Big]dt + \omega^{-1}\sum_{l=1}^{K}\mathbb{E}[\kappa_{l,v}\beta_{l,v}^{takeup}].$$

where

$$
\begin{aligned}
\omega &\coloneqq \sum_{l=1}^{K}\int_0^\infty \mathbb{E}\Big[\tau_{l,v}(t)\Big]dt + \sum_{l=1}^{K}\mathbb{E}[\kappa_{l,v}] \\
\tau_{l,v}(t) &\coloneqq Var\big(\mathbb{1}\{T_l > t\}|v\big)\Big(\mathbb{E}[\widetilde{W}_k|l,v,T_l > t] - \mathbb{E}[\widetilde{W}_k|l,v,T_k < t]\Big) \\
\kappa_{k,v} &\coloneqq Var(D_l|v)\big(\mathbb{E}[\widetilde{W}_k|l,v] - \mathbb{E}[\widetilde{W}_k]\big)\mathbb{E}[T_k|k,v]
\end{aligned}
$$

Proof. By Lemma 5, β^{2SLS} is equivalently given by the estimand for the regression of Y on

$\widehat{W}$. It follows then from the Frisch-Waugh-Lovell Theorem, that β_j^{2SLS} is the estimand given by first regressing $\widehat{W}_j$ on all other elements of $\widehat{W}$ and then regressing Y on the residual of this regression, denoted as $\widetilde{W}_j$.

It follows that if Z satisfies Assumptions 1, 2, and 3, then so does $\widetilde{W}_j$ as $\widetilde{W}_j$ is a linear combination of Z. Thus, replacing W with $\widetilde{W}_j$ in Lemmas 5, 6, and 7 and Theorem 1 gives the result. □

Corollary 8. *Suppose that Assumptions 1, 2, and 3 hold and Assumption 4 holds for treatment k. Then denote $\widetilde{D}_k = \widetilde{W}_k$ and $\tilde{T}_k = \widetilde{W}_{2k}$, respectively. Then:*

$$\kappa_{k,v} > 0$$
$$\tau_{2k,v}(t) > 0$$

Proof. Continuous Case:
Under Assumptions 1, 2, and 3, the weight on $\beta_{k,v}^{ACRT}(t_0)$ at an arbitrary t_0 in the support of T_k is given by

$$\tau_{k,v}(t_0) \coloneqq Var\big(\mathbb{1}\{T_k > t\}|v\big)\Big(\mathbb{E}[\widetilde{T}_k|k,v,T_k > t] - \mathbb{E}[\widetilde{T}_k|k,v,T_k \leq t]\Big).$$

Now consider three mutually exclusive and exhaustive subcases:

a) $Var\big(\mathbb{1}\{T_k > t\}|v\big) > 0$

b) $Var\big(\mathbb{1}\{T_k > t\}|v\big) = 0$

Continuous Subcase (a):
We have $Var\big(\mathbb{1}\{T_k > t_0\}|v\big) > 0$. Then consider:

$$\Big(\mathbb{E}[\widetilde{T}_k|k,v,T_k > t_0] - \mathbb{E}[\widetilde{T}_k|k,v,T_k \leq t_0]\Big).$$

By Assumption 4, we have that $\mathbb{E}[\widetilde{T}_k|k,v,T_k > t_0] - \mathbb{E}[\widetilde{T}_k|k,v,T_k < t_0] \geq 0$. It follows that $\tau_{k,v}(t_0) \geq 0$.

Continuous Subcase (b):
We have that $Var\big(\mathbb{1}\{T_k > t_0\}|v\big) = 0$, which ensures that $\tau_{k,v}(t_0) = 0$.

Thus, Assumptions 1, 2, 3 and 4 are sufficient for $\tau_{k,v}(t) \geq 0$. To see that they are necessary, suppose that $\tau_{k,v}(t_0) < 0$. Then this is only possible in Subcase (a). In particular, it must be that

$$\mathbb{E}[\widetilde{T}_k\ \big|k,v,T_k > t_0] < \mathbb{E}[\widetilde{T}_k\ \big|k,v,T_k \leq t_0].$$

So Assumption 4 cannot hold.

Discrete Case: Under Assumptions 1, 2, and 3, the weight $\kappa_{k,v}$ for response type v is given by

$$\kappa_{k,v} \coloneqq (\mathbb{E}[\widetilde{D}_k|k,v] - \mathbb{E}[\widetilde{D}_k])Var(D_k|v)\mathbb{E}[T_k|k,v]$$

Consider the two mutually exclusive and exhaustive subcases:

(a) $Var(D_k|v) > 0$

(b) $Var(D_k|v) = 0$

Discrete Subcase (a):
We have that $Var(D_k|v) > 0$. Then $\kappa_{k,v} > 0$ if and only if

$$\frac{\mathbb{E}[\widetilde{D}_k|k,v] - \mathbb{E}[\widetilde{D}_k]}{\mathbb{E}[T_k|k,v]} \geq 0$$

This occurs if and only if the numerator and denominator have the same sign or the numerator is zero. Thus, Assumption 4 is necessary and sufficient for $\kappa_{k,v} > 0$ in Discrete Subcase (a).

Discrete Subcase (b):
We have that $Var(D_k|v) = 0$. Thus, $\kappa_{k,v} = 0$ for any value of $\mathbb{E}[\widetilde{D}_k|k,v]$. □

Corollary 9. *Under Assumptions 1, 2, 3, and 5, the 2SLS estimand for the effect of W_k is given by*

$$\begin{aligned} (W_k = T_k) \rightarrow \quad \beta_k^{2SLS} &= \frac{\int_0^\infty \mathbb{E}[\tau_{k,v}(t)\beta_{k,v}^{ACRT}(t)]dt}{\int_0^\infty \mathbb{E}[\tau_{k,v}(t)]dt} \\ (W_k = D_k) \rightarrow \quad \beta_k^{2SLS} &= \frac{\mathbb{E}[\kappa_{k,v}\beta_{k,v}^{takeup}]}{\mathbb{E}[\kappa_{k,v}]} \end{aligned}$$

Proof. Under Assumptions 1, 2, and 3 the weights on $\beta_{j,v}^{takeup}$ and $\beta_{j,v}^{ACRT}$ for type v are given by $\kappa_{j,v}$ and $\tau_{j,v}(t)$, respectively, where

$$\begin{aligned} \tau_{j,v}(t) &:= Var\big(\mathbb{1}\{T_j > t\}|v\big)\Big(\mathbb{E}[\widetilde{W}_k|j, v, T_j > t] - \mathbb{E}[\widetilde{W}_k|k, v, T_j \le t]\Big) \\ \kappa_{j,v} &:= (\mathbb{E}[\widetilde{W}_k|j, v] - \mathbb{E}[\widetilde{W}_k])Var(D_j|v)\mathbb{E}[T_j|j, v] \end{aligned}$$

Discrete Case

We have three subcases for when $\kappa_{j,v} = 0$:

(a) $Var(D_j|v) = 0$

(b) $\mathbb{E}[T_j|j, v] = 0$

(c) $\mathbb{E}[\widetilde{W}_k|j, v] = \mathbb{E}[\widetilde{W}_k]$

We want to show that mean independence of $\widetilde{W}_k$ from D_j is necessary and sufficient to ensure subcase (c) when subcases (a) and (b) do not hold. To see this, first notice that Assumption 1 implies

$$\mathbb{E}[\widetilde{W}_k|v] = \mathbb{E}[\widetilde{W}_k].$$

Furthermore, by definition, $\widetilde{W}_k$ is mean independent of D_j conditional on v, if and only if

$$\mathbb{E}[\widetilde{W}_k|j, v] = \mathbb{E}[\widetilde{W}_k|v].$$

So we have that $\widetilde{W}_k$ is mean independent of D_j if and only if

$$\mathbb{E}[\widetilde{W}_k|j, v] = \mathbb{E}[\widetilde{W}_k|v] = \mathbb{E}[\widetilde{W}_k].$$

Continuous Case

We have three subcases for when $\tau_{j,v}(t) = 0$ for any arbitrary t in the support of T_j:

(a) $Var\big(\mathbb{1}\{T_j > t\}|v\big) = 0$

(b) $\mathbb{E}[\widetilde{W}_k|j, v, T_j > t] = \mathbb{E}[\widetilde{W}_k|k, v, T_j \le t]$

Thus, we must show that mean independence of $\widetilde{W}_k$ from T_j is necessary and sufficient to ensure subcase (b) when subcase (a) does not hold.

(*Mean Independence → (b)*):

First, show that mean independence implies subcase (b). To that end, consider any t_0 in the support of T_j. Then note that by Assumption 5, for any t' in the support of T_j such that $t' < t_0$, it holds that

$$\mathbb{E}[\widetilde{W}_k|j, v, T_j = t'] = \mathbb{E}[\widetilde{W}_k|j, v].$$

This implies that

$$\mathbb{E}[\widetilde{W}_k|j, v, T_j \le t_0] = \mathbb{E}[\widetilde{W}_k|j, v].$$

By a similar argument, it holds that

$$\mathbb{E}[\widetilde{W}_k|j,v,T_j > t_0] = \mathbb{E}[\widetilde{W}_k|j,v].$$

(*Subcase (b) → Mean Independence*):
To show that subcase (b) implies mean independence, we will first show that subcase (b) implies that the conditional mean of $\widetilde{W}_k$ is constant over any arbitrary interval that overlaps with the support of T_j. Then we will show that this implies that $\widetilde{W}_k$ is point-wise mean independent of T_j.
To that end, assume that subcase (b) holds. Then for all t in the support of T_j, we have that

$$\mathbb{E}[\widetilde{W}_k|j,v,T_j > t] = \mathbb{E}[\widetilde{W}_k|j,v,T_j \leq t].$$

Now consider the conditional expectation $\mathbb{E}[\widetilde{W}_k|j,v]$, which can be expressed as:

$$\mathbb{E}[\widetilde{W}_k|j,v] = \mathbb{P}(T_j > t|j,v)\ \mathbb{E}[\widetilde{W}_k|j,v,T_j > t] + \mathbb{P}(T_j \leq t|j,v)\ \mathbb{E}[\widetilde{W}_k|j,v,T_j \leq t]$$

Substituting with $\mathbb{E}[\widetilde{W}_k|j,v,T_j > t] = \mathbb{E}[\widetilde{W}_k|j,v,T_j \leq t]$ gives

$$\begin{aligned}\mathbb{E}[\widetilde{W}_k|j,v] &= (1 - \mathbb{P}(T_j \leq t|j,v))\ \mathbb{E}[\widetilde{W}_k|j,v,T_j \leq t] + \mathbb{P}(T_j \leq t|j,v)\ \mathbb{E}[\widetilde{W}_k|j,v,T_j \leq t]\\ &= \mathbb{E}[\widetilde{W}_k|j,v,T_j \leq t]\\ &= \mathbb{E}[\widetilde{W}_k|j,v,T_j > t].\end{aligned}$$

Now consider any t_l, t_u such that $t_l < t_u$ and there exists at least one $t_0 \in [t_l, t_u]$ such that t_0 is within the support of T_j. Then by the above we have that

$$\mathbb{E}[\widetilde{W}_k|j,v,T_j \leq t_l] = \mathbb{E}[\widetilde{W}_k|j,v,T_j > t_u] = \mathbb{E}[\widetilde{W}_k|j,v]. \tag{8}$$

By the Law of Iterated Expectations, we have that

$$\begin{aligned}\mathbb{E}[\widetilde{W}_k|j,v] = \mathbb{P}(T_j \leq t_l|j,v)\mathbb{E}[\widetilde{W}_k|j,v,T_j \leq t_l] &+ \mathbb{P}(t_l < T_j \leq t_u|j,h)\mathbb{E}[\widetilde{W}_k|j,v,t_l < T_j \leq t_u]\\ &+ \mathbb{P}(T_j > t_u|j,v)\mathbb{E}[\widetilde{W}_k|j,v,T_j > t_u].\end{aligned}$$

Substituting according to Equation (8) gives

$$\begin{aligned}\mathbb{E}[\widetilde{W}_k|j,v] = \mathbb{P}(T_j \leq t_l|j,v)\mathbb{E}[\widetilde{W}_k|j,v] &+ \mathbb{P}(t_l < T_j \leq t_u|j,v)\mathbb{E}[\widetilde{W}_k|j,v,t_l < T_j \leq t_u]\\ &+ \mathbb{P}(T_j > t_u|j,v)\mathbb{E}[\widetilde{W}_k|j,v].\end{aligned}$$

This implies that

$$\mathbb{P}(t_l < T_j \leq t_u|j,v)\mathbb{E}[\widetilde{W}_k|j,v] = \mathbb{P}(t_l < T_j \leq t_u|j,v)\mathbb{E}[\widetilde{W}_k|j,v,t_l < T_j \leq t_u]$$

and hence

$$\mathbb{E}[\widetilde{W}_k|j,v] = \mathbb{E}[\widetilde{W}_k|j,v,t_l < T_j \leq t_u].$$

Thus, we have shown that the conditional mean of $\widetilde{W}_k$ is constant across any arbitrary interval in $\mathbb{R}$ overlapping with the support of T_j. It follows that

$$\mathbb{E}[\widetilde{W}_k|j, v, t_l < T_j \leq t_0] = \mathbb{E}[\widetilde{W}_k|j, v].$$

Now consider taking the limit as $t_l \to t_0$. Because t_l was chosen arbitrarily, the right-hand side remains constant. Thus we have that

$$\lim_{t_l \to t_0} \mathbb{E}[\widetilde{W}_k|j, v, t_l < T_j \leq t_0] = \mathbb{E}[\widetilde{W}_k|j, v, T_j = t_0] = \mathbb{E}[\widetilde{W}_k|j, v].$$

Because t_0 was chosen to be an arbitrary element in the support of T_j, this implies that $\mathbb{E}[\widetilde{W}_k|j, v]$ is mean independent of T_j.
Joint Case
We have shown in the above that under Assumptions 1, 2, and 3, $\mathbb{E}[\widetilde{W}_k|j, v]$ is mean independent of T_j, if and only if $\tau_{j,v}(t) = 0$ for all t in the support of T_j and that $\mathbb{E}[\widetilde{W}_k|v]$ is mean independent of D_j if and only if $\kappa_{j,v} = 0$. That is

$$\begin{aligned} \mathbb{E}[\widetilde{W}_k|j, v, T_j] = \mathbb{E}[\widetilde{W}_k|j, v] &\quad \to \quad \tau_{j,v}(t) = 0 \\ \mathbb{E}[\widetilde{W}_k|j, v] = \mathbb{E}[\widetilde{W}_k] &\quad \to \quad \kappa_{j,v} = 0 \end{aligned}$$

It follows that if $\mathbb{E}[\widetilde{W}_k|v]$ is mean independent of all W_j where $W_j \neq W_k$, then we have by definition of mean independence that

$$\mathbb{E}[\widetilde{W}_k|v, W] = \mathbb{E}[\widetilde{W}_k|v, W_k],$$

where we may note that in the case where $W_k = T_k$ that necessarily we have that

$$\begin{aligned} \mathbb{E}[\widetilde{W}_k|D_k = 1, v, T_k \neq 0] = \mathbb{E}[\widetilde{W}_k|v, T_k \neq 0] \\ \mathbb{E}[\widetilde{W}_k|D_k = 0, v, T_k = 0] = \mathbb{E}[\widetilde{W}_k|v, T_k = 0] \end{aligned}$$

because $\mathbb{P}(D_k = 0, T_k \neq 0) = \mathbb{P}(D_k = 1, T_k = 0) = 0$.
Furthermore, this implies that

$$\begin{aligned} (W_k = T_k) \to \quad \beta_k^{2SLS} &= \frac{\int_0^\infty \mathbb{E}[\tau_{k,v}(t)\beta_{k,v}^{ACRT}(t)]dt}{\int_0^\infty \mathbb{E}[\tau_{k,v}(t)]dt} \\ (W_k = D_k) \to \quad \beta_k^{2SLS} &= \frac{\mathbb{E}[\kappa_{k,v}\beta_{k,v}^{takeup}]}{\mathbb{E}[\kappa_{k,v}]} \end{aligned}$$

□

Theorem 3. *Under Assumptions 1, 2, 3, 4, and 5, the 2SLS coefficient on W_k, given by β_k^{2SLS}, represents a weighted average of $\beta_{k,v}^{ACRT}(t)$ if $W_k = T_k$ or a weighted average of $\beta_{k,v}^{takeup}$ if $W_k = D_k$. Specifically, the weighted averages are given by*

$$\beta_k^{2SLS} = \frac{\int_0^\infty \mathbb{E}[\tau_{k,v}(t)\beta_{k,v}^{ACRT}(t)]dt}{\int_0^\infty \mathbb{E}[\tau_{k,v}(t)]dt} \qquad (W_k = T_k)$$

$$\beta_k^{2SLS} = \frac{\mathbb{E}[\kappa_{k,v}\beta_{k,v}^{takeup}]}{\mathbb{E}[\kappa_{k,v}]} \qquad (W_k = D_k)$$

where the weights are positively valued for all v and given by

$$\begin{aligned}\tau_{k,v}(t) &:= Var\big(\mathbb{1}\{T_k > t\}|v\big)\Big(\mathbb{E}[\widetilde{W}_k|k,v,T_k > t] - \mathbb{E}[\widetilde{W}_k|k,v,T_k < t]\Big)\\ \kappa_{k,v} &:= Var(D_k|v)\big(\mathbb{E}[\widetilde{W}_k|k,v] - \mathbb{E}[\widetilde{W}_k]\big)\mathbb{E}[T_k|k,v]\end{aligned}$$

Proof. Under Assumptions 1, 2, 3, 4, and 5, Corollary 9 shows that, β_k^{2SLS} is given by

$$\begin{aligned}\beta_k^{2SLS} &= \frac{\int_0^\infty \mathbb{E}[\tau_{k,v}(t)\beta_{k,v}^{ACRT}(t)]dt}{\int_0^\infty \mathbb{E}[\tau_{k,v}(t)]dt} && (W_k = T_k)\\ \beta_k^{2SLS} &= \frac{\mathbb{E}[\kappa_{k,v}\beta_{k,v}^{takeup}]}{\mathbb{E}[\kappa_{k,v}]} && (W_k = D_k)\end{aligned}$$

where

$$\begin{aligned}\tau_{k,v}(t) &:= Var\big(\mathbb{1}\{T_k > t\}|v\big)\Big(\mathbb{E}[\widetilde{W}_j|k,v,T_k > t] - \mathbb{E}[\widetilde{W}_j|k,v,T_k < t]\Big)\\ \kappa_{k,h} &:= Var(D_k|v)\big(\mathbb{E}[\widetilde{W}_k|k,v] - \mathbb{E}[\widetilde{W}_k]\big)\mathbb{E}[T_k|k,v].\end{aligned}$$

By Corrollary 8, if Assumption 4 also holds, then for all h, $\tau_{k,v}(t) > 0$ for all t in the support of T_k if $W_k = T_k$, or if $W_k = D_k$ then $\kappa_{k,v} > 0$. □

Corollary 10. *Let $W = (D_1, \ldots, D_K, T_1, \ldots, T_k)$ and define W_{-l} as the vector including all elements of W except for some element W_l. Now suppose that Z satisfies Assumptions 1, 2, and 3 but is correlated with W_{-l}.*

$$\beta^{2SLS} = \boldsymbol{\omega}^{-1}\sum_{k=1}^K \int_0^\infty \mathbb{E}[\tau_{k,v}(t)\beta_{k,v}^{ACRT}(t)]dt + \boldsymbol{\omega}^{-1}\sum_{k=1}^K \mathbb{E}\Big[\kappa_{k,v}\beta_{k,v}^{takeup}\Big]$$

where

$$\begin{aligned}\widehat{W}_{-l} &= Var(Z)^{-1}Cov(Z, W_{-l})\\ \boldsymbol{\omega} &:= \sum_{k=1}^K \int_0^\infty \mathbb{E}[\lambda_{k,v}(t)]dt + \sum_{k=1}^K \mathbb{E}[\kappa_{k,v}]\\ \tau_{k,v}(t) &:= Var(\mathbb{1}\{T_k > t\}|v)\Big(\mathbb{E}[\widehat{W}_{-l}|k,v,T_k > t] - \mathbb{E}[\widehat{W}_{-l}|k,v,T_k \leq t]\Big)\\ \kappa_{k,v} &:= Var(D_k|v)\Big(\mathbb{E}[\widehat{W}_{-l}|k,v] - \mathbb{E}[\widehat{W}_{-l}]\Big)\mathbb{E}[T_k|k,v]\end{aligned}$$

Proof. The results of Theorems 1 and 2 still hold because the only element of β^{2SLS} that depends on W_{-l} is the predicted value of $\widehat{W}$. All other elements were derived by relying on the unobserved first stage response function h for the true treatment vector W, potential outcomes g over the true treatment vector W, and the by treating $\widehat{W}$ as a function of Z. Thus, the result holds by substituting $\widehat{W}$ for $\widehat{W}_{-l}$, given by

$$\widehat{W}_{-l} = Var(Z)^{-1}Cov(Z, W_{-l}).$$

□

C Comparing 2SLS Estimand to OLS

Other recent econometric identification results have revealed that the OLS estimator can be represented as a weighted average of treatment effects, similar to those derived in the main analysis of this paper. In particular, Callaway et al. (2024) show that the difference-in-differences estimator with a (single) continuous treatment can be expressed as either a weighted average of causal response parameters or as an average takeup effect (in their terminology, a level effect). In their analysis, they consider the OLS estimator for a single continuous treatment under the "parallel trends" assumption. However, their results can be broadened to any general OLS estimator with a continuous treatment under the assumption that treatment takeup is exogenous to the outcome (for a difference-in-differences estimator, the outcome is first demeaned).

For comparison, I first replicate these results here for a single continuous treatment in order to provide a consistent notation for comparison across estimators and then show that the 2SLS esitmator can be built in a mirroring way. First, I provide identification for the OLS estimator as a weighted average of treatment effects under endogeneity. This provides a greater intuition into the tradeoffs associated with 2SLS estimation in comparison to OLS, as well as the role of various exogeneity assumptions often employed in regression-based research designs. The results below show that 2SLS can be viewed as first constructing group average effects for each response type. These effects serve as the building blocks for the overall estimator and remove endogeneity by marginalizing out the influence of the second stage residual component U. As shown throughout the main text of this paper, this approach still faces the risk that the 2SLS weights may introduce other issues, in particular they may be negative or introduce cross-contamination.

To that end, let the $W_i = (1, D_i, T_i)$ be a vector of the constant and assignment to a continuously-valued treatment for individual i. As before, D_i indicates whether the individual receives the treatment at all (takeup) and T_i indicates the level of the treatment received (intensity). I again assume that potential outcomes are given by the function $Y = g(W, U)$, where U is multidimensional and thus allows for unconstrained heterogeneity in effects. Important is that (W, U) are assumed to be jointly distributed and so W is endogenously assigned. The OLS estimator is given by the following regression equation

$$Y_i = \beta W_i + U_i \tag{9}$$

The identification result for the OLS estimand from Equation (9) is given by the following theorem.

Theorem 11. *The OLS estimand of β from Equation (9) is given as*

$$\beta^{OLS} = \frac{1}{Var(W)}\Big(\int_{-\infty}^{\infty} \beta^{ACRT}(t)\tau(t)dt + \kappa\beta^{takeup} + Cov\big(g(0,U), \mathbb{E}[W|D,U]\big)\Big) + \\ \frac{1}{Var(W)}\Big(\underbrace{\int_{-\infty}^{\infty} \mathbb{E}\Big[\tau_u(t)\Delta\beta_u^{ACRT}\Big]dt + \mathbb{E}[\kappa_u \Delta\beta_u^{takeup}]}_{selection\ bias}\Big)$$

where

$$\begin{aligned}
\tau_u(t) &= \mathbb{P}(T > t|U)(\mathbb{E}[W|D = 1, U, T > t] - \mathbb{E}[W|D = 1, U]) \\
\kappa_u &= \mathbb{P}(D = 1|U)(\mathbb{E}[W|D = 1, U] - \mathbb{E}[W]) \\
\tau(t) &= \mathbb{E}[\tau_u(t)] \\
\kappa &= \mathbb{E}[\kappa_u]
\end{aligned}$$

Proof. The OLS estimator is given by

$$Y_i = \beta W_i + u_i$$

where $W = (1, D_i, T_i)$. The OLS estimand is given by

$$\beta^{OLS} = \frac{Cov(Y, W)}{Var(W)}$$

Proceeding as before, we will first focus on $Cov(Y, W)$. By the Law of Total Covariance, this is equivalently given by

$$Cov\big(g(T, U), W\big) = \mathbb{E}\Big[Cov\big(g(T, U), W\ |D, U\big)\Big] + Cov\big(\mathbb{E}[g(T, U)|D, U], \mathbb{E}[W|D, U]\big)$$

Now focus on the first covariance term

$$\begin{aligned}
\mathbb{E}\Big[Cov\big(g(T, U), W\ |D, U\big)\Big] &= \mathbb{E}\Big[\mathbb{P}(D = 1|U)\mathbb{E}\Big[(W - \mathbb{E}[W|D = 1, U]) \int_{-\infty}^{\infty} g'(t, U)\mathbb{1}\{T > t\}dt\ \Big|D = 1, U\Big]\Big] \\
&= \int_{-\infty}^{\infty} \mathbb{E}\Big[\mathbb{P}(D = 1|U)g'(t, U)\mathbb{E}\Big[(W - \mathbb{E}[W|D = 1, U])\mathbb{1}\{T > t\}\ \Big|D = 1, U\Big]\Big]dt \\
&= \int_{-\infty}^{\infty} \mathbb{E}\Big[\Big(g'(t, U) - \beta^{ACRT}(t) + \beta^{ACRT}(t)\Big)\tau_u(t)\Big]dt \\
&= \int_{-\infty}^{\infty} \mathbb{E}\Big[\Big(g'(t, U) - \beta^{ACRT}(t)\Big)\tau_u(t) + \beta^{ACRT}(t)\tau_u(t)\Big]dt \\
&= \int_{-\infty}^{\infty} \mathbb{E}\Big[\Big(g'(t, U) - \beta^{ACRT}(t)\Big)\tau_u(t) + \beta^{ACRT}(t)\tau_u(t)\Big]dt \\
&= \int_{-\infty}^{\infty} \mathbb{E}\Big[\tau_u(t)\Delta\beta_u^{ACRT}\Big]dt + \int_{-\infty}^{\infty} \beta^{ACRT}(t)\mathbb{E}\Big[\tau_u(t)\Big]dt \\
&= \underbrace{\int_{-\infty}^{\infty} \mathbb{E}\Big[\tau_u(t)\Delta\beta_u^{ACRT}\Big]dt}_{selection\ bias} + \int_{-\infty}^{\infty} \beta^{ACRT}(t)\tau(t)dt.
\end{aligned}$$

The first equality holds because when $D = 0$, the vector W is just a constant and the following terms are defined as

$$\begin{aligned}
\tau_u(t) &= \mathbb{P}(T > t|U)(\mathbb{E}[W|D = 1, U, T > t] - \mathbb{E}[W|D = 1, U]) \\
\tau(t) &= \mathbb{E}[\tau_u(t)] \\
\Delta\beta_u^{ACRT} &= g'(t, U) - \beta^{ACRT}(t).
\end{aligned}$$

Before turning to the second covariance term, define $D_0 := \mathbb{1}\{(D = 0)\}$ and $D_1 := \mathbb{1}\{(D = 1)\}$.

Then we have that

$$\begin{aligned}
Cov\big(\mathbb{E}[Y|D,U],\mathbb{E}[W|D,U]\big) &= Cov\big(\mathbb{E}[YD_0+YD_1|D,U],\mathbb{E}[W|D,U]\big)\\
&= Cov\big(\mathbb{E}[YD_0|D,U],\mathbb{E}[W|D,U]\big)+Cov\big(\mathbb{E}[YD_1|D,U],\mathbb{E}[W|D,U]\big)
\end{aligned}$$

Now, consider each individual covariance in turn. First, consider $Cov\big(\mathbb{E}[YD_1|D,U],\mathbb{E}[W|D,U]\big)$. Then, we may note that $\mathbb{E}[YD_1|D]$ is given as

$$\begin{aligned}
\mathbb{E}[Y|D=1,u] &= \mathbb{E}[Y|D=1,u]-\mathbb{E}[Y|D=0,u]+\mathbb{E}[Y|D=0,u]\\
&= \beta_u^{takeup}+\mathbb{E}[Y|D=0,u]\\
&= \beta_u^{takeup}+(\mathbb{E}[Y|D=1]-\mathbb{E}[Y|D=0])-(\mathbb{E}[Y|D=1]-\mathbb{E}[Y|D=0])+\mathbb{E}[Y|D=0,u]\\
&= \beta_u^{takeup}+\beta^{takeup}-\beta^{takeup}+\mathbb{E}[Y|D=0,u]\\
&= \beta^{takeup}+(\beta_u^{takeup}-\beta^{takeup})+\mathbb{E}[Y|D=0,u]\\
&= \beta^{takeup}+\Delta\beta_u^{takeup}+\mathbb{E}[Y|D=0,u].
\end{aligned}$$

where β^{takeup}, β_u^{takeup}, and $\Delta\beta_u^{takeup}$ are defined as

$$\begin{aligned}
\beta^{takeup} &:= \mathbb{E}[Y|D=1]-\mathbb{E}[Y|D=0]\\
\beta_u^{takeup} &:= \mathbb{E}[Y|D=1,u]-\mathbb{E}[Y|D=0,u]\\
\Delta\beta_u^{takeup} &:= \beta_u^{takeup}-\beta^{takeup}.
\end{aligned}$$

$$\begin{aligned}
&Cov\big(\mathbb{E}[YD_1|D,U],\mathbb{E}[W|D,U]\big)\\
&\quad= \sum_u \mathbb{P}(D=1,u)\mathbb{E}[Y|D=1,u]\Big(\mathbb{E}[W|D=1,u]-\mathbb{E}[W]\Big)\\
&\quad= \sum_u \mathbb{P}(D=1,u)\Big(\beta^{takeup}+\Delta\beta_u^{takeup}+\mathbb{E}[Y|D=0,u]\Big)\Big(\mathbb{E}[W|D=1,u]-\mathbb{E}[W]\Big)\\
&\quad= \kappa\beta^{takeup}+\sum_u \mathbb{P}(D=1,u)\Big(\Delta\beta_u^{takeup}+\mathbb{E}[Y|D=0,u]\Big)\Big(\mathbb{E}[W|D=1,u]-\mathbb{E}[W]\Big)
\end{aligned}$$

where

$$\kappa := Var(D)\Big(\mathbb{E}[W|D=1]-\mathbb{E}[W|D=0]\Big).$$

Now consider $Cov\big(\mathbb{E}[YD_0|D,U],\mathbb{E}[W|D,U]\big)$. This is equivalently given as

$$Cov\big(\mathbb{E}[YD_0|D,U],\mathbb{E}[W|D,U]\big)=\sum_u \mathbb{P}(D=0,u)\mathbb{E}[Y|D=0,u]\Big(\mathbb{E}[W|D=0,u]-\mathbb{E}[W]\Big).$$

Returning to the covariance of conditional expectations, we have

$$\begin{aligned}
Cov\big(\mathbb{E}[Y|D,U],\mathbb{E}[W|D,U]\big) &= Cov\big(\mathbb{E}[YD_0|D,U],\mathbb{E}[W|D,U]\big)+Cov\big(\mathbb{E}[YD_1|D,U],\mathbb{E}[W|D,U]\big)\\
&= \kappa\beta^{takeup}+\mathbb{E}[\kappa_u\Delta\beta_u^{takeup}]+Cov\Big(\mathbb{E}[Y|D=0,U],\mathbb{E}[W|D,U]\Big)\\
&= \kappa\beta^{takeup}+\underbrace{\mathbb{E}[\kappa_u\Delta\beta_u^{takeup}]}_{selection\ bias}+Cov\Big(g(0,U),\mathbb{E}[W|D,U]\Big)
\end{aligned}$$

where $\kappa_u \coloneqq \mathbb{P}(D=1|u)(\mathbb{E}[W|D=1,u]-\mathbb{E}[W])$, noting that $\kappa = \mathbb{E}[\kappa_u]$. □

The above result shows that OLS estimates the average takeup effect, the ACRT, and the conditional baseline mean of the outcome, subject to two selection biases. The selection biases reflect heterogeneity in treatment effects across the joint distribution of treatment assignment W and U. The first selection bias is captured by the integral of $\mathbb{E}[\tau_u(t)\Delta\beta_u^{ACRT}]$ across the support of T. This generally reflects how selection occurs along the intensive margin. The second selection bias is captured by the expectation $\mathbb{E}\Big[\kappa_u \Delta\beta_u^{takeup}\Big]$. This captures how selection along the extensive margin into treatment is correlated with differences in takeup effects.

The weights can also be viewed as aggregating heterogeneity in treatment effects. Because the terms $\Delta\beta_u^{takeup}$ and $\Delta\beta_u^{ACRT}$ are necessarily mean zero, a non-zero selection bias indicates that the weighted average over-represents some individuals and under-represents others on the basis of unobserved characteristics u.

This also provides the intuition in how endogeneity biases the OLS estimator through the above derivation. The misrepresentation of heterogeneous effects captures variation in responses driven by the joint variation across U and W. From this perspective, it is not clear whether the OLS selection bias terms are driven by variation in treatment assignment or variation in unobserved factors captured by U.

The following theorem provides the corresponding derivation for 2SLS and helps provide intuition on how 2SLS addresses this selection bias.

Theorem 1'. *Under Assumptions 1, 2, 3, 4, and 5, the 2SLS estimand, denoted by by β^{2SLS}, can be decomposed into the average treatment effect and a weighted average of the heterogeneity in treatment effects across response types V*

$$\beta^{2SLS} = Var(W)^{-1}\sum_{k=1}^{K}\Big(\int_{-\infty}^{\infty}\widehat{\tau}_k(t)\beta_k^{ACRT}(t)dt + \widehat{\kappa}_k\beta^{takeup} + \mathbb{E}\Big[\widehat{\kappa}_{k,v}\mathbb{E}[Y|D_k=0,v]\Big]\Big)$$

$$+ Var(W)^{-1}\sum_{k=1}^{K}\Big(\underbrace{\int_{-\infty}^{\infty}\mathbb{E}\Big[\widehat{\tau}_{k,v}(t)\Delta\beta_{k,v}^{ACRT}(t)\Big]dt + \mathbb{E}\Big[\widehat{\kappa}_{k,v}\Delta\beta_{k,v}^{takeup}\Big]}_{treatment\ effect\ heterogeneity}\Big)$$

where the following terms are defined as

$$\begin{aligned}
\widehat{\tau}_{k,v}(t) &\coloneqq \mathbb{P}(T_k > t|v)\Big(\mathbb{E}[\widehat{W}|k,v,T_k>t] - \mathbb{E}[\widehat{W}|k,v]\Big) \\
\widehat{\tau}_k(t) &\coloneqq \mathbb{E}[\widehat{\tau}_{k,v}(t)] \\
\widehat{\kappa}_{k,v} &\coloneqq \mathbb{P}(D_k=1|v)(\mathbb{E}[\widehat{W}|k,v] - \mathbb{E}[\widehat{W}]) \\
\widehat{\kappa}_k &\coloneqq \mathbb{E}[\widehat{\kappa}_{k,v}] \\
\beta_{k|\widehat{\tau}}^{ACRT}(t) &\coloneqq \mathbb{E}[\beta_{k,v}^{ACRT}(t)|\widehat{\tau}_{k,v}(t) \neq 0] \\
\Delta\beta_{k,v}^{ACRT}(t) &\coloneqq \beta_{k,v}^{ACRT}(t) - \beta_{k,v|\widehat{\tau}}^{ACRT}(t) \\
\beta_{k|\widehat{\kappa}}^{takeup} &\coloneqq \mathbb{E}[Y|D_k=1,\widehat{\kappa}_{k,v}\neq 0] - \mathbb{E}[Y|D_k=0,\widehat{\kappa}_{k,v}\neq 0] \\
\beta_{k,v}^{takeup} &\coloneqq \mathbb{E}[Y|D_k=1,v] - \mathbb{E}[Y|D_k=0,v] \\
\Delta\beta_{k,v}^{takeup} &\coloneqq \beta_{k,v}^{takeup} - \beta_{k,v|\widehat{\kappa}}^{takeup}.
\end{aligned}$$

Proof. From Lemma 5, the 2SLS estimator is given by

$$\beta^{2SLS} = Var(\widehat{W})^{-1} \sum_{k=1}^{K} Cov(\widehat{W}, YD_k)$$

where by the Law of Total Covariance

$$Cov(\widehat{W}, YD_k) = \sum_{k=1}^{K} \mathbb{E}[Cov(\widehat{W}, YD_k|D_k, V)] + \sum_{k=1}^{K} Cov\Big(\mathbb{E}[\widehat{W}|D_k, V], \mathbb{E}[YD_k|D_k, V]\Big).$$

First, from formula (7) in Lemma 6, we have that

$$Cov(\widehat{W}, YD_k|k, v) = \int_{-\infty}^{\infty} \beta_{k,v}^{ACRT}(t)\mathbb{P}(T_k > t|k, v)\Big(\mathbb{E}[\widehat{W}|k, v, T_k > t] - \mathbb{E}[\widehat{W}|k, v]\Big)dt.$$

Then taking the expectation of $Cov(\widehat{W}, YD_k|D_k, V)$ gives

$$\begin{aligned}
\mathbb{E}[Cov(\widehat{W}, YD_k|D_k, V)] &= \sum_{v} \mathbb{P}(D_k = 1, v)Cov(\widehat{W}, Y|k, v) \\
&= \sum_{v} \mathbb{P}(v)\mathbb{P}(D_k = 1|v)Cov(\widehat{W}, Y|k, v) \\
&= \sum_{v} \mathbb{P}(v) \int_{-\infty}^{\infty} \widehat{\tau}_{k,v}(t)\beta_{k,v}^{ACRT}(t)dt \\
&= \int_{-\infty}^{\infty} \mathbb{E}\Big[\widehat{\tau}_{k,v}(t)\beta_{k,v}^{ACRT}(t)\Big]dt \\
&= \int_{-\infty}^{\infty} \mathbb{E}\Big[\widehat{\tau}_{k,v}(t)\big(\beta_{k|\widehat{\tau}}^{ACRT}(t) + \Delta\beta_{k,v}^{ACRT}(t)\big)\Big]dt \\
&= \int_{-\infty}^{\infty} \widehat{\tau}_{k}(t)\beta_{k|\widehat{\tau}}^{ACRT}(t)dt + \int_{-\infty}^{\infty} \mathbb{E}\Big[\widehat{\tau}_{k,v}(t)\Delta\beta_{k,v}^{ACRT}(t)\Big]dt
\end{aligned}$$

where the following terms are defined as

$$\begin{aligned}
\widehat{\tau}_{k,v}(t) &:= \mathbb{P}(T_k > t|v)\Big(\mathbb{E}[\widehat{W}|k, v, T_k > t] - \mathbb{E}[\widehat{W}|k, v]\Big) \\
\widehat{\tau}_{k}(t) &:= \mathbb{E}[\widehat{\tau}_{k,v}(t)] \\
\beta_{k|\widehat{\tau}}^{ACRT}(t) &:= \mathbb{E}[\beta_{k,v}^{ACRT}(t)|\widehat{\tau}_{k,v}(t) \neq 0] \\
\Delta\beta_{k,v}^{ACRT}(t) &:= \beta_{k,v}^{ACRT}(t) - \beta_{k}^{ACRT}(t)
\end{aligned}$$

Now consider $Cov(\mathbb{E}[\widehat{W}|D_k, V], \mathbb{E}[YD_k|D_k, V])$. From Lemma 7, this is given by

$$Cov\Big(\mathbb{E}[\widehat{W}|D_k, v], \mathbb{E}[YD_k|D_k, v]\Big) = \mathbb{E}\Big[\widehat{\kappa}_{k,v}\mathbb{E}[Y|k, V]\Big].$$

where $\widehat{\kappa}_{k,v} := \mathbb{P}(D_k = 1|v)(\mathbb{E}[\widehat{W}|k, v] - \mathbb{E}[\widehat{W}])$. For a particular v, we may rewrite $\mathbb{E}[Y|k, v]$ as follows

$$\begin{aligned}\mathbb{E}[Y|k,v] &= \mathbb{E}[Y|D_k=1,v] - \mathbb{E}[Y|D_k=0,v] + \mathbb{E}[Y|D_k=0,v]\\ &= \beta^{takeup}_{k,v} + \mathbb{E}[Y|D_k=0,v]\\ &= \beta^{takeup}_{k,v} + \beta^{takeup}_{k|\widehat{\kappa}} - \beta^{takeup}_{k|\widehat{\kappa}} + \mathbb{E}[Y|D_k=0,v]\\ &= \beta^{takeup}_{k|\widehat{\kappa}} + \Delta\beta^{takeup}_{k,v} + \mathbb{E}[Y|D_k=0,v]\end{aligned}$$

where the following terms are defined as

$$\begin{aligned}\beta^{takeup}_{k|\widehat{\kappa}} &:= \mathbb{E}[Y|D_k=1,\widehat{\kappa}_v \neq 0] - \mathbb{E}[Y|D_k=0,\widehat{\kappa}_v \neq 0]\\ \beta^{takeup}_{k,v} &:= \mathbb{E}[Y|D_k=1,v] - \mathbb{E}[Y|D_k=0,v]\\ \Delta\beta^{takeup}_{k,v} &:= \beta^{takeup}_{k,v} - \beta^{takeup}_{k|\widehat{\kappa}}.\end{aligned}$$

Thus, we have that

$$\begin{aligned}&Cov\Big(\mathbb{E}[\widehat{W}|D_k,v],\mathbb{E}[g^k(T_k,U)D_k|D_k,v]\Big)\\ &\qquad= \mathbb{E}\Big[\widehat{\kappa}_{k,v}\mathbb{E}[Y|k,V]\Big]\\ &\qquad= \mathbb{E}\Big[\widehat{\kappa}_{k,v}\Big(\beta^{takeup}_{k,\widehat{\kappa}} + \Delta\beta^{takeup}_{k,v} + \mathbb{E}[Y|D_k=0,v]\Big)\Big]\\ &\qquad= \widehat{\kappa}_k\beta^{takeup}_{k,\widehat{\kappa}} + \mathbb{E}\Big[\widehat{\kappa}_{k,v}\Delta\beta^{takeup}_{k,v}\Big] + \mathbb{E}\Big[\widehat{\kappa}_{k,v}\mathbb{E}[Y|D_k=0,v]\Big].\end{aligned}$$

with $\widehat{\kappa}_k = \mathbb{E}[\widehat{\kappa}_{k,v}]$.

□

The above result shows that 2SLS can be represented in a mirroring form to the OLS estimator, except that 2SLS constructs weights with $\widehat{W}$ and V, instead of W and U, and that the targeted average effects are local to those response types for whom the instrument induces variation in treatment assignment. This is not surprising as the second stage point estimates are equivalently obtained from a regression of Y on $\widehat{W}$. However, with 2SLS heterogeneity is captured only across response types V. The weighted average of treatment effect heterogeneity, however, is still prone to being misrepresented because an unweighted average would be mean zero, but in general the weighted heterogeneity terms do not sum to zero. This leaves a question as to in what way the 2SLS estimator marks an improvement over the OLS estimator, given that both estimators are prone to misrepresenting treatment effect heterogeneity?

One view to this is that the 2SLS estimators can be viewed as first estimating group average treatment effects within each response type v using the exogenous variation induced by the instrument. These group effects provide exogenous building blocks to aggregate up. From this perspective, the group average heterogeneity marginalizes out the heterogeneity across U so that the remaining heterogeneity in treatment effects is only across W. 2SLS still misrepresents this heterogeneity, but the influence of U is at least removed.

D Threshold Crossing Model

Threshold crossing models are encompassed within the general framework described in Section 4. These models generally are employed to estimate marginal treatment effects following the approach of Heckman and Vytlacil (2005), which are more readily interpretable in comparison to the general weighted LATE. Given the relative pervasiveness of these results in the literature, I include a brief discussion relating my results to the above setup here, but for a more complete discussion direct readers to related studies like Arteaga (2023), Bhuller and Sigstad (2023), Chyn et al. (2024), and Humphries et al. (2024). As well, I note that threshold crossing models are generally considered within the framework of discretely measured treatments. My discussion here does not intend to map those identification results to the multi-treatment continuous setting. Rather, I map the model framework and estimating assumptions to the multi-treatment continuous setting, which serves to highlight that the assumptions place empirically strong restrictions on judge sentencing behavior.

A simple variation of a threshold crossing model often employed in the literature can be obtained by assuming that judges hold common preferences over offenses α_j and common beliefs Φ_j, that Φ_j is rank invariant in v with respect to w, and that preferences over punishments R_j are commonly ordered. This is summarized below:

$$\begin{array}{rcll}
\Phi_j(c|q_w, w, v_i) & = & \Phi_{j'}(c|q_w, w, v_i) & \forall\,(j, j') \\
\alpha_j & = & \alpha & \forall j \\
\Phi_j(c|q_w, w, v_i) > \Phi_j(c|q_w, w, v_i') & \Leftrightarrow & \Phi_j(c|q_{w'}, w', v_i) > \Phi_j(c|q_{w'}, w', v_i') & \forall\,(w, w', v_i, v_i') \\
R_j(w) > R_j(w') & \Leftrightarrow & R_{j'}(w) > R_{j'}(w') & \forall\,(j, j', w, w')
\end{array}$$

The restrictions on Φ_j and α_j induces a single index for measuring the severity of defendants that is common to all judges and hence variation in sentencing only arises from differences in stringency. The intuition of the above model is that all judges share a common ordering of punishments W and a common ordering of defendants i in terms of their perceived severity. Furthermore, these orderings are independent of each other and defined across the support of v and so the ranking of an individual is invariant to the choice of W[12]. Hence, we can express an individual i's indexed ranking along a single index $\phi_i \in [0, 1]$.

The above model is attractive and is often applied in related studies as it implies "pairwise monotonicity", which is defined as

Assumption 6. *(Pairwise Monotonicity) For any pair of judges j, j', where w_{ji} and $w_{j'i}$ represents the punishment given by judge j to defendant i, the following holds:*

$$R_j(w_{ji}) \geq R_{j'}(w_{j'i})\ \forall i \quad OR \quad R_j(w_{ji}) \leq R_{j'}(w_{j'i})\ \forall i$$

Pairwise monotonicity is a stronger form of the average conditional monotonicity assumption because it effectively means that judge punishments can also be ordered. However, whether the

[12] As disussed in Humphries et al. (2024), plea deals give a situation where the independence of defendant and punishment orderings are likely violated. In the context of the above model, this occurs when the cost function R is allowed to depend on whether a defendant accepts a plea. For clarity, let the punishment be given as $(W, Plea)$. Then the distribution $(W, Plea = 0, V)$ is likely very different to $(W, Plea = 1, V)$ because some individuals will never be offered the plea and some will never accept. As a consequence, Φ_j under $Plea = 1$ is not defined for some v. This can lead to violations in monotonicity as weaker punishments are accepted through plea deals by selectively more severe defendants.

assumption and overall model hold is in actuality an empirical matter. The identification of marginal treatment effects lastly relies on the assumption that judges share a common support for punishments W, which can be stated as

Assumption 7. *(Common support of W) $\mathbb{P}(W = w|j) > 0$ for all judges j and $w \in Supp(W)$*

Assumption 7 implies that for every judge j and punishment $w \in Supp(W)$, there exists a $v_i \in Supp(V)$ such that $w_{ij} = w$. This allows for stronger identification results because discrete jumps from probation to incarceration across a judge's preferences always occur at the same local point in $Supp(W)$. This establishes a clear comparison region for the marginally incarcerated defendant. For example, if the maximum possible probation sentence is 5 years, then the above set of assumptions would suggest that the marginally incarcerated defendant under judge j would receive a probation sentence of 5 years under a slightly more lenient judge j'. Under this assumption, we may express the first stage as a monotonically increase function of ϕ_i for each judge j, $h_j(\phi)$.

The standard estimator for the marginal treatment effect of prison takes the form of a Wald estimator

$$\beta^{MTE}(p) = \lim_{p' \to p} \frac{\mathbb{E}[Y_i|\mathbb{P}(prison|Z) = p] - \mathbb{E}[Y_i|\mathbb{P}(prison|Z) = p']}{p - p'}$$

$\beta^{MTE}(p)$ has the form a univariate discrete 2SLS estimator and thus the results of Section 5 apply. As a consequence, even under the restrictions of the above model, if W is multidimensional and continuous then $\beta^{MTE}(p)$ generally captures the effect across all elements of W. The above model does provide some structure around this, however. In particular, convergence in the two alternative probabilities of incarceration (p, p') generally reflects convergence in the distribution of $h_j(\phi)$ as p' converges to p. Thus, even local IV estimators made in the style of Heckman and Vytlacil (2005) will generally reflect the full effect of these distributional differences. Recent studies have addressed by adopting a

it is clear from the results of Section 5 that $\beta^{MTE}(p)$ will not generally reflect

The combination of the above assumptions is emprically very strong. First, the exclusion restriction requires that if judge decisions on whether to incarcerate are correlated with judge decisions on the length of incarceration (or probation), that one must also include sentence lengths in W. Pairwise monotonicity requires then that a relatively strict judge is strict across every dimension of W, while the common support assumption on W then requires that judges do not have gaps in their individual sentencing patterns. Gaps in sentencing patterns might occur if, for example, some judges are willing to assign defendants to probation up to a maximum of three years before turning to incarceration while other judges are only willing to assign probation up to a maximum of two years.

E Testing Robustness of Condtional Exclusion Restriction when Omitting Sentence Lengths from Endogenous Measures

Here I assess the performance of alternative research designs to controlling for omitted treatments. In each case, I keep the first stage equations the same as in the tests above to ensure that differences in the results are not driven by differences in the prediction for the included treatments. Because omission of correlated treatments is likely to induce an exclusion restriction violation, I focus on tests for the conditional exclusion restriction in which I regress all elements of W (those specified in Equation (3)) on a subset of the first stage predictions. The two alternative approaches I take are the following:

1. Include only the first stage predictions for the extensive margin (being granted probation and receiving a fine) while excluding all other elements of W (Table B5)

$$w_{k,i} = \tilde{\alpha}_1 \widehat{Probation}_i + \tilde{\alpha}_2 \widehat{Fine}_i + \boldsymbol{\theta X}_i + \eta_{zip} + \gamma_r + u_i. \tag{10}$$

2. Repeat Test 1 but include instruments for omitted treatments as controls, denoted by $Z^{exclude}$ (Table B6)

$$w_{k,i} = \tilde{\alpha}_1 \widehat{Probation}_i + \tilde{\alpha}_2 \widehat{Fine}_i + \pi Z_i^{exlcude} + \boldsymbol{\theta X}_i + \eta_{zip} + \gamma_r + u_i. \tag{11}$$

For each case, the test strongly rejects the null hypothesis and thus approach 2 above does not appear to adequately address omitted treatment biases. Furthermore, there is not a clear causal interpretation of the results under the omitted treatment bias. For example, one cannot interpret 2SLS estimands as "collapsing" the sentence lengths in. For one, the effects of omitted treatments for some response types will necessarily be negatively weighted. This removes any interpretation of averaging. Two, it is unclear where the omitted treatment biases appear. In approach 1 and 2 above, the omitted treatments might enter into the estimand for probation or fine and it is not possible to assess the extent to which this occurs for either measure. Three, the effects will generally represent different subpopulations. This last point relates to the definition of takeup effects and ACRT parameters. Specifically, takeup effects only reflect the effect for those who possibly receive, e.g., probation or not. An "always taker" of prison would necessarily drop out of the takeup effect. However, an "always taker" of prison would not be excluded from ACRT effects. This provides a clear example as to why takeup and ACRT effects cannot be interpretted jointly because both are local to overlapping but different regions of the distribution of (V, W). In this sense, the ACRT does not (solely) represent the effect of extensive margin compliers.

Table B5: Wald Test for the Conditional Exclusion Restriction Excluding Continuous Sentencing Measures

	Probation		Prison Time		Probation Time		Stayed Time	
	$X = 0$	$X = 1$	$X = 0$	$X = 1$	$X = 0$	$X = 1$	$X = 0$	$X = 1$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Multi-unit Residence	0.7183	0.8040	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**
1{Unit Area > Median}	0.5688	0.5499	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**
1{Unit Value > Median}	0.2971	0.3766	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**
Family residence	0.8123	0.9212	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**
1{Age > Median}	0.6594	0.6384	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**
Male	**0.0073**	0.1927	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**
White	0.2617	**0.0212**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**
Other Race	0.8058	0.1057	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**
Divorced last year	0.9121	0.6012	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**
Paternity claim last year	0.8719	0.4294	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**
Evicted last year	0.7359	0.4747	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**
Committed crime last year	0.5807	0.2241	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**

Notes: Results from Wald tests for the conditional exclusion restriction when continuous sentencing measures are omitted from the second stage. Rows describe the variable used to determine the sample split, while columns indicate the sub-sample tested on for each sentencing dimension. Each entry indicates a p-value from an F-test that proper weights are satisfied. Bolded entries indicate results that are statistically significant after adjusting for a false disovery rate assumed to be equal to α.

Table B6: Wald Test for the Conditional Exclusion Restriction Instruments Controls for Continuous Sentencing Measures

	Probation		Prison Time		Probation Time		Stayed Time	
	$X = 0$	$X = 1$	$X = 0$	$X = 1$	$X = 0$	$X = 1$	$X = 0$	$X = 1$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Multi-unit Residence	0.7597	0.8272	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**
1{Unit Area > Median}	0.5792	0.5550	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**
1{Unit Value > Median}	0.4509	0.5681	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**
Family residence	0.9285	0.9920	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0144**	**0.0000**
1{Age > Median}	0.6343	0.5905	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**
Male	**0.0101**	0.2326	**0.0000**	**0.0000**	**0.0000**	**0.0000**	0.1648	**0.0000**
White	0.2166	**0.0113**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**
Other Race	0.8328	0.2385	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0069**
Divorced last year	0.9110	0.6092	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0772**
Paternity claim last year	0.8578	0.4242	**0.0000**	**0.0001**	**0.0000**	**0.0000**	**0.0000**	0.4044
Evicted last year	0.8005	0.5696	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0001**
Committed crime last year	0.6932	0.4570	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0001**

Notes: Results from Wald tests for the conditional exclusion restriction when omitting sentence lengths from the second stage endogenous variables but including controls for average judge sentence lengths. Rows describe the variable used to determine the sample split, while columns indicate the sub-sample tested on for each sentencing dimension. Each entry indicates a p-value from an F-test that proper weights are satisfied. Bolded entries indicate results that are statistically significant after adjusting for a false disovery rate assumed to be equal to α.

F Robustness to Inclusion of Non-probation and Non-incarcerated Defendants

Table B7: Cross-sectional Effects of Courtroom Sentencing on Crime (All Cases)
5 Years Post-sentencing

	All Crimes		Property Crimes		Violent Crimes		Other Crimes	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Crimes by Defendant								
Extensive Margin								
Not Incarcerated	0.0856	0.7629^{**}	0.0268	0.2388^{*}	−0.0407	0.2599	0.0995^{*}	0.2642
	(0.0730)	(0.3513)	(0.0457)	(0.1360)	(0.0255)	(0.1631)	(0.0585)	(0.2216)
Fine	-0.1663^{***}	−0.0922	-0.0626^{*}	−0.0324	0.0041	0.0343	−0.1078	−0.0942
	(0.0600)	(0.0871)	(0.0321)	(0.0334)	(0.0216)	(0.0313)	(0.0646)	(0.0766)
Prison Durations								
Prison Years		0.0257		0.0076		0.0228		−0.0048
		(0.0396)		(0.0164)		(0.0223)		(0.0188)
$<$ 5 Years		-0.3548^{**}		-0.1211^{*}		−0.1430		−0.0907
		(0.1618)		(0.0631)		(0.0869)		(0.0999)
Probation Durations								
Probation Years		-0.3278^{**}		-0.1750^{**}		−0.0724		−0.0804
		(0.1561)		(0.0752)		(0.0828)		(0.1079)
Stayed Years		0.0081		0.0160		0.0460		−0.0538
		(0.1086)		(0.0579)		(0.0531)		(0.0635)
Panel B: Crimes by Others on Property								
Extensive Margin								
Not Incarcerated	-0.1241^{*}	0.5544	0.0064	0.1945	−0.0370	0.2232^{*}	-0.0935^{*}	0.1367
	(0.0681)	(0.3786)	(0.0294)	(0.1532)	(0.0243)	(0.1207)	(0.0487)	(0.2837)
Fine	0.0079	0.0005	0.0036	0.0244	0.0136	0.0132	−0.0092	−0.0372
	(0.0583)	(0.0713)	(0.0218)	(0.0252)	(0.0179)	(0.0220)	(0.0457)	(0.0519)
Prison Durations								
Prison Years		0.0543		0.0199		0.0066		0.0278
		(0.0470)		(0.0154)		(0.0144)		(0.0313)
$<$ 5 Years		-0.3253^{*}		−0.1019		-0.0989^{*}		−0.1245
		(0.1812)		(0.0650)		(0.0565)		(0.1307)
Probation Durations								
Probation Years		0.3585^{*}		−0.0769		0.1537^{**}		0.2817^{**}
		(0.1815)		(0.0814)		(0.0652)		(0.1086)
Stayed Years		-0.2811^{**}		0.0244		−0.0710		-0.2345^{***}
		(0.1301)		(0.0629)		(0.0445)		(0.0645)
N	29,330	29,330	29,330	29,330	29,330	29,330	29,330	29,330
Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Pre-trial Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Continuous Effects	No	Yes	No	Yes	No	Yes	No	Yes

Notes: 2SLS estimates for the effect of sentencing on criminal outcomes in the 60 months post-sentencing when including guilty defendants who are not sentenced to prison or probation. Each column is a 2SLS regression of the outcome on sentencing dimensions, instrumented with assigned judge's expected sentence. Panel A presents estimates for crimes commited by the defendant. Panel B presents estimates for crime rates committed by others (excluding the defendant) living at the same residence as the defendant as a rate per property unit on the address. Odd columns exclude continuous dimensions of court sentences from the estimating equation while even columns include continuous sentencing dimensions. Standard errors clustered by judge in parentheses. *** = p $<$ 0.01, ** = p $<$ 0.05, * = p $<$ 0.1.

Table B8: Panel Estimates for the Effects of Sentencing Releases on Crime (All Cases) 5 Years Post-sentencing

	All Crimes	Property Crimes	Violent Crimes	Other Crimes
	(1)	(2)	(3)	(4)
Panel A: Crimes by Defendant				
Extensive Margin				
Released (Prison)	0.1584**	0.0590**	0.0435*	0.0559
	(0.0651)	(0.0281)	(0.0251)	(0.0378)
Released (Probation)	−0.3711**	−0.0819	−0.0835	−0.2056**
	(0.1618)	(0.0706)	(0.0565)	(0.1008)
Probation	0.2933**	0.1110**	0.0369	0.1454**
	(0.1173)	(0.0519)	(0.0378)	(0.0643)
Fine	−0.0174	−0.0081	0.0077	−0.0169
	(0.0161)	(0.0068)	(0.0063)	(0.0145)
Prison Durations				
Release × Years Prison	−0.0886	0.0026	−0.0367	−0.0545
	(0.0756)	(0.0328)	(0.0279)	(0.0440)
Probation Durations				
Release × Years Probation	0.2382***	0.0431	0.0622*	0.1329**
	(0.0854)	(0.0387)	(0.0321)	(0.0539)
Stayed Sentence (Years)	−0.0635*	−0.0166	−0.0044	−0.0424**
	(0.0342)	(0.0144)	(0.0112)	(0.0190)
Panel B: Crimes by Others on Property				
Extensive Margin				
Released (Prison)	0.1287**	0.0286	0.0663***	0.0339
	(0.0507)	(0.0278)	(0.0235)	(0.0405)
Released (Probation)	−0.1253	−0.0454	−0.0464	−0.0335
	(0.1003)	(0.0567)	(0.0420)	(0.0765)
Probation	0.1647**	0.0361	0.0763**	0.0523
	(0.0760)	(0.0476)	(0.0315)	(0.0658)
Fine	0.0070	0.0057	0.0087	−0.0075
	(0.0153)	(0.0064)	(0.0053)	(0.0092)
Prison Durations				
Release × Years Prison	−0.0992**	−0.0309	−0.0431*	−0.0252
	(0.0489)	(0.0274)	(0.0238)	(0.0346)
Probation Durations				
Release × Years Probation	0.0769	0.0282	0.0252	0.0236
	(0.0579)	(0.0313)	(0.0238)	(0.0417)
Stayed Sentence (Years)	−0.0491**	−0.0055	−0.0111	−0.0324**
	(0.0228)	(0.0132)	(0.0103)	(0.0126)
N	146,745	146,745	146,745	146,745
Fixed Effects	Yes	Yes	Yes	Yes
Pre-trial Controls	Yes	Yes	Yes	Yes

Notes: 2SLS estimates corresponding to Equation (4) for the effect of sentencing on criminal outcomes in the year of the observation when including guilty defendants who are not sentenced to probation or prison. Each column is a 2SLS regression of the outcome on sentencing dimensions, instrumented with the assigned judge's expected sentence. Panel A presents estimates for crimes commited by the defendant. Panel B presents estimates for crime rates committed by others (excluding the defendant) living at the same residence as the defendant as a rate per property unit on the address. Standard errors clustered by judge in parentheses. *** = $p < 0.01$, ** = $p < 0.05$, * = $p < 0.1$.