University of Toronto
Department of Economics

Working Paper 782

A Sharp Test for the Judge Leniency Design

By Mohamed Coulibaly, Yu-Chin Hsu, Ismael Mourifie and
Yuanyuan Wan

August 06, 2024

A Sharp Test for the Judge Leniency Design*

Mohamed Coulibaly† Yu-Chin Hsu‡ Ismael Mourifié§ Yuanyuan Wan¶

July 14, 2024

Abstract

We propose a sharp test to assess the validity of the exclusion and monotonicity assumption in judge leniency designs. Our sharp test exploits all the relevant information in the observed data distribution to refute the model and will not make discordant recommendations. When the validity of the design is rejected, we show that a variant of the marginal treatment effect (MTE) can be identified under weaker partial monotonicity and partial exclusion assumptions. We apply our test to the Philadelphia court data studied by Stevenson (2018) and demonstrate that it outperforms existing non-sharp tests by significant margins in simulation studies.

Keywords: Judge Leniency Design, Instrumental Variables, Specification Test, Moment Inequalities.

JEL Classification: C12, C14, C21 and C26

*We have greatly benefited from insightful comments from Bocar Ba. We also thank William Arbour, Kory Kroft, and Brigham Frandsen for their valuable discussions. Mourifié and Wan thank the support from the SSHRC Insight Grant #435190500. Hsu gratefully acknowledges the research support from the National Science and Technology Council of Taiwan (NSTC112-2628-H-001-001), the Academia Sinica Investigator Award of the Academia Sinica, Taiwan (AS-IA-110-H01), and from the Center for Research in Econometric Theory and Applications of National Taiwan University (Grant No. 113L8601).

†Department of Applied Economics, HEC Montréal. Email: mohamed.coulibaly@hec.ca.

‡Institute of Economics, Academia Sinica; Department of Finance, National Central University; Department of Economics, National Chengchi University; CRETA, National Taiwan University. E-mail: ychsu@econ.sinica.edu.tw.

§Corresponding author. Department of Economics, Washington University in St. Louis, One Brookings Drive St. Louis, MO 63130-4899, USA. E-mail: ismaelm@wustl.edu.

¶Department of Economics, University of Toronto. E-mail: yuanyuan.wan@utoronto.ca.

1 Introduction

We propose a novel sharp test to assess the validity of the judge leniency design, which has emerged as a prominent instrumental variable (IV) approach in recent years, particularly in empirical research exploring causal effects within the criminal justice system. This design has proven beneficial in investigating the impacts of various interactions with the legal system, like pretrial detentions and incarcerations, on subsequent outcomes such as recidivism rates, conviction probabilities, and employment prospects. What sets the judge leniency design apart is its distinctive feature of randomly assigning judges to different cases, with each judge handling a significant number of cases while having discretion over the final decision. The random assignment of judges enhances the credibility of this IV approach and has led to its increasing popularity among researchers (Kling, 2006; Di Tella and Schargrodsky, 2013; Aizer and Doyle Jr, 2015; Mueller-Smith, 2015).[1] Importantly, the judge leniency design's random assignment feature extends beyond the context of criminal justice, making it a valuable methodology in diverse research contexts, including medicine, patents and startups, bankruptcy protection, evictions, and access to foster care (see Doyle Jr, Graves, Gruber, and Kleiner, 2015; Farre-Mensa, Hegde, and Ljungqvist, 2020; Dobbie, Goldsmith-Pinkham, and Yang, 2017; Gross and Baron, 2022).[2]

However, in addition to the random assignment, an instrumental variable must adhere to two additional crucial conditions: (i) an exclusion restriction, which means that judges' actions should only influence the treatment and should not have any direct influence on the defendant's future outcomes; and (ii) a monotonicity restriction, which means that judges should consistently exhibit more or less leniency. For example, if a defendant is treated (detained) by a lenient judge, she should always be treated (detained) by a less lenient judge. While the assumption of random assignment is typically satisfied in the judge leniency design, the validity of the exclusion and monotonicity assumptions can be questionable. In practice, trial decisions (treatment) are often multidimensional, including incarceration, fines, community service, sentence length and others (Johnson, 2014).

[1] Kling (2006) exploits randomized judge assignment along with judge propensities to instrument for incarceration length, aiming to investigate the causal impact of incarceration on labor market outcomes.

[2] For example, Doyle Jr, Graves, Gruber, and Kleiner (2015) employs the judge leniency design in the medical context to examine the impact of ambulance companies on patients in emergencies, relying on the pseudo-random assignment of ambulance companies to patients. Similarly, Dobbie, Goldsmith-Pinkham, and Yang (2017) uses the leniency of randomly assigned bankruptcy judges as an instrument to study the implications of Chapter 13 bankruptcy protection on future financial events.

These decisions impact the future outcomes. Because different judges may have varying attitudes on these decisions, the exclusion restriction can be violated if some of the decisions are unobserved or uncontrolled. Furthermore, Abrams, Bertrand, and Mullainathan (2012) and Stevenson (2018) argue there is considerable heterogeneity in how judges rank defendants when considering various types of offences. If this heterogeneity is not observed, then it is possible that judges exhibit varying levels of leniency under different circumstances, and the monotonicity assumption would be violated. Therefore, offering a statistical test to evaluate the validity of the judge leniency design becomes a highly relevant empirical question.

In this paper, we revisit the set of testable implications of the marginal treatment effect framework derived in Heckman and Vytlacil (2005), which are essentially many inequality restrictions applied to the joint distribution of observed outcomes, treatment status, and propensity scores. We show that a specific tractable subset of these inequality restrictions characterizes the sharp testable implications of the judge leniency design. These sharp testable implications possess the unique quality of exploiting all available information within the data distribution that is useful to refute the validity of the judge leniency design. We propose asymptotically valid and consistent semi-nonparametric and semiparametric tests based on these tractable testable implications.

Numerous efforts have been made to test the judge leniency design in the existing literature. A common approach involves providing separate evidence for the validity of the individual assumptions made in the judge leniency design. For instance, to assess the random assignment of judges, Dobbie, Grönqvist, Niknami, Palme, and Priks (2018) examine whether a measure of judge stringency (the instrumental variable) correlates with baseline cases and family characteristics of criminal defendants. Regarding the monotonicity assumption, they test an implication that requires the first-stage estimates to be non-negative for all subsamples. Bhuller, Dahl, Løken, and Mogstad (2018) and Norris, Pecenco, and Weaver (2021) employ similar individual testing approaches. Assessing the assumptions individually is effective in empirical scenarios where researchers know which assumption to test and have prior knowledge that other assumptions hold. However, when no such prior knowledge is available, testing assumptions individually can fail to screen out violations. In fact, the three key assumptions may collectively impose certain constraints on the observable data-generating process (DGP), which could not be detected by examining only the testable implications of each assumption in isolation.

Unlike individually testing each assumption, Frandsen, Lefgren, and Leslie (2023) propose a joint test for all assumptions underlying the judge leniency design. Their test leverages the property that, in the judge leniency design, the average outcome at the judge level should exhibit a smooth relationship with the propensity score (or the judge-level treatment probability). It ought to have a bounded slope, where the bounds depend on the limits of the outcome variable's support. Although Frandsen, Lefgren, and Leslie (2023)'s testable implication has the desirable property that it assesses all the assumptions simultaneously, we show there is still relevant information in the data distribution essential for evaluating the judge leniency design's validity, but not used in Frandsen, Lefgren, and Leslie (2023)'s testable implication. This difference is also demonstrated by the simulation and empirical studies reported in Section 4.

To the best of our knowledge, our test is the only sharp test available for assessing the validity of the judge leniency design. In other words, our testable implications exhaust all the information in the observed data distribution. As seen in previous methods, non-sharp tests have practical virtue when there is no easily tractable characterization of the sharp testable implications of a model's assumptions. If a non-sharp rejects, it conveys an informative result that the assumptions should be rejected. However, there are also important trade-offs to consider. First, a non-sharp test can have no power against certain violations since it does not consider all possible constraints on the data distribution. Second, different non-sharp tests can lead to discordant empirical results and potentially misleading interpretations of the estimand of interest (see Kédagni, Li, and Mourifié, 2020). For instance, two different non-sharp tests may produce conflicting results because they consider different aspects of the observed data distribution. Our sharp test addresses both issues as it is a consistent test built upon sharp testable implications and, therefore, a useful complement to the existing tests.

As a potential alternative to the existing non-sharp tests, one may consider testing the validity of the judge leniency design employing some of the existing sharp tests developed for the Local Average Treat Effect (LATE) framework, i.e., Kitagawa (2015), Huber and Mellace (2015), and Mourifié and Wan (2017). However, it is worth noting these tests may over-reject since they are based on a priori direction in the monotonicity assumption and are not directly applicable in the context of judge leniency design. For instance, in the judge leniency design, the number of judges can be quite large, and in some cases, it might even be infinite, especially when judges' types are continuous. In such scenarios, the number of potential directions to consider becomes large, possibly infinite. Imposing a specific ex-ante direction in the judge leniency design is therefore overly restrictive, and to consider all possible directions might be impractical or impossible. Furthermore, imposing an incorrect a priori direction bears an additional risk of model misspecification. These issues highlight the need for a more flexible testing approach, like the one proposed in this paper, which is free from making overly restrictive assumptions on the direction of monotonicity.

Finally, while our test is mainly motivated by testing judge leniency designs, it can be applied to test the identifying assumptions in a general Marginal Treatment Effect framework with continuous or discrete instrument variables, which has been applied to various empirical settings. See Carneiro, Heckman, and Vytlacil (2011); Kowalski (2016); Brinch, Mogstad, and Wiswall (2017), among many others. In the context of judge leniency designs, this also means that our test does not require observing a judge's identity and accommodates continuous judge types.

We organize the rest of the paper as follows. Section 2 presents the analytical framework and the sharp testable implications of the judge leniency design. Section 3 presents the testing procedures. In Section 4, we show the results of the simulations and discuss our empirical illustration. In Section 5, we explore approaches to salvage the judge leniency design when its sharp testable implications are violated. The last section concludes the paper, and the proofs are collected in the online supplementary materials.

2 Model and Sharp Testable Implications

We adopt the potential outcomes framework. Let the observed treatment indicator be $D \in \{0, 1\}$. For example, in the judge leniency design, $D = 1$ denotes incarceration. We denote by Z the judge's observable type and let $Z \in \mathcal{Z} \subseteq \mathbb{R}^{d_z}$ for $d_z \geq 1$. $Y_d(z) \in \mathcal{Y} \subseteq \mathbb{R}$ denotes the potential outcome of interests (e.g. recidivism) when the treatment and the judge's type are externally set to $D = d$, and $Z = z$, respectively. Similarly, $D(z)$ denotes the potential treatment when the judge's type is externally set to $Z = z$. Let $Y = Y_1(Z)D + Y_0(Z)(1 - D)$ be the observed outcome. For the moment, we omit observed defendant and case covariates X (such as time and courtroom of the trial) for ease of notation. The identification analysis in this section can be extended by conditioning on X. We will also discuss the implementation of our test in the presence of X in Section 3.2.

In our setting, Z can be multidimensional; it can also be continuous, discrete, or mixed. For example, if there is a group of judges $\mathcal{J}$, and their identities are observed, then $Z \in \mathcal{J}$ can be chosen as the identity of the judge assigned to the defendant. This is the instrumental variable that Frandsen, Lefgren, and Leslie (2023, FLL hereafter) consider. On the other hand, we allow scenarios in which the judge's identity is unobserved but with observed characteristics. In this case, Z may contain a set of continuous or discrete variables, such as the judge's experience, gender, and race. The literature mainly relies on the following assumptions to evaluate the causal effects of treatment D on outcome Y.

Assumption 2.1 (Random assignment of judges) $Z \perp (Y_0(z), Y_1(z), D(z); z \in \mathcal{Z})$.

Assumption 2.2 (Exclusion restriction) *There is no direct effect of judges' type on the potential outcomes. For $d \in \{0, 1\}$, $Y_d(z) = Y_d$ for all $z \in \mathcal{Z}$.*

Assumption 2.3 (Monotonicity) *For any pair $(z, z') \in \mathcal{Z} \times \mathcal{Z}$ either $D(z) \geq D(z')$ for all defendants or $D(z) \leq D(z')$ for all defendants.*

A particular feature of the judge leniency design is judges are usually randomly assigned to different cases, making the random assignment assumption likely to hold in practice. Hence, we assume that Assumption 2.1 holds throughout the paper. However, Assumptions 2.2 and 2.3 are usually less credible. Assumption 2.2 means the effect of judges on the potential outcomes must necessarily transit through their effect on treatment assignment. Assumption 2.3 requires any defendants treated (incarcerated) by a more lenient judge be also treated if assigned to a less lenient one. Heckman and Vytlacil (2005) refer to the monotonicity assumption as a uniformity condition since it imposes a restriction across judges rather than the shape of a function for a particular judge. Vytlacil (2002) provides an equivalent characterization of the monotonicity assumption, which can be stated as follows:

Assumption 2.4 (Single Threshold-Crossing: STC) *The judge treatment assignment mechanism is governed by the following threshold crossing model $D = 1\{\nu(Z) \geq U\}$ for some measurable and non-trivial function ν, where the distribution of U is absolutely continuous.*

Under Assumptions 2.1 and 2.4, we can rewrite the threshold crossing model without loss of generality as follows:

$$D = 1\left\{F_U(\nu(Z)) \geq F_U(U)\right\} \equiv 1\left\{P(Z) \geq V\right\},$$

where $F_U(\cdot)$ is the distribution function of U, $P(\cdot) \equiv F_U(\nu(\cdot))$ is identified from the observed variables (D, Z) by $P(z) = \mathbb{P}(D = 1|Z = z)$, and $V \equiv F_U(U) \sim Uniform[0, 1]$. Hereafter, we will write $P(Z)$ as P when it causes no confusion. Let $\mathcal{P} \subseteq [0, 1]$ denote the support of $P(Z)$. It is worth noting that the STC does not impose a priori direction in z in the monotonicity condition since Assumption 2.4 is equivalent to Assumption 2.3 (Vytlacil, 2002). Under Assumptions 2.1, 2.2 and 2.4, the judge leniency design model can be equivalently written as:

$$\begin{aligned} Y &= Y_1 D + Y_0(1 - D), && (2.1)\\ D &= 1\left\{P(Z) \geq V\right\}. && (2.2) \end{aligned}$$

Assumptions 2.1, 2.2 and 2.4 (equivalently Assumptions 2.1 to 2.3) impose some restrictions on the joint distribution of the observed variables $(Y, D, P(Z))$, which we will characterize in Theorem 1. But before stating the theorem, let us give the intuition of the testable implications. Let $g : \mathcal{Y} \to \mathbb{R}^+$ be a nonnegative real integrable function such that $\mathbb{E}|g(Y_d)| < \infty$. Taking $d = 0$ as an illustration. For any pair $(p, p') \in \mathcal{P} \times \mathcal{P}$ such that $p \leq p'$, we have:

$$\begin{aligned} \mathbb{E}[g(Y)(1 - D)|P = p] &= \mathbb{E}[g(Y_0)1\{V \geq P\}|P = p] = \mathbb{E}[g(Y_0)1\{V \geq p\}] \\ &\geq \mathbb{E}[g(Y_0)1\{V \geq p'\}] = \mathbb{E}[g(Y_0)1\{V \geq P\}|P = p'] = \mathbb{E}[g(Y)(1 - D)|P = p']. \end{aligned}$$

The first and fourth equalities hold by Assumption 2.4 (STC) and Assumption 2.2 (exclusion); the second and third equalities hold because of Assumption 2.1 (random assignment), and the inequality holds because $p \leq p'$. Intuitively, under the assumptions of the judge leniency design, if a defendant is released by judge p', then he/she would necessarily be released by judge p since judge p is more lenient than judge p'. On the other hand, there can exist a set of defendants who were released by a type p judge, but not by a type p' judge: a group of “compliers”. Because $g(Y_0)$ is nonnegative, the average $g(Y_0)$

for this group of compliers is also nonnegative, delivering the inequality we see from the displayed equation above. The discussion is formalized in the following theorem.

Theorem 1 (Sharp characterization of the Judges' IV design assumptions) *Let the collection of variables $(Y, D, Y_1, Y_0, P(Z))$ define a potential outcome model $Y = Y_1 D + Y_0(1 - D)$.*

(i) If Assumptions 2.1, 2.2 and 2.4 (equivalently Assumptions 2.1 to 2.3) hold, then for all $y, y' \in \mathcal{Y}$, $\mathbb{P}(y < Y \leq y', D = 1|P = p)$ and $-\mathbb{P}(y < Y \leq y', D = 0|P = p)$ are non-decreasing in p for all $p \in \mathcal{P}$.

(ii) If for all $y, y' \in \mathcal{Y}$, $\mathbb{P}(y < Y \leq y', D = 1|P = p)$ and $-\mathbb{P}(y < Y \leq y', D = 0|P = p)$ are non-decreasing in p for all $p \in \mathcal{P}$, there exists a joint distribution of $(\tilde{V}, \tilde{Y}_1, \tilde{Y}_0, P(Z))$ such that Assumptions 2.1, 2.2 and 2.4 hold, and $(\tilde{Y}, \tilde{D}, P(Z))$ has the same distribution as $(Y, D, P(Z))$.

The proof of Theorem 1 is collected in the supplementary materials Section B. The testable implications in Theorem 1(i) are a subset of the implications previously derived in Heckman and Vytlacil (2005, Appendix A), who show for any non-negative integrable function, i.e. $g(\cdot) : \mathcal{Y} \to \mathbb{R}^+$, $\mathbb{E}[g(Y)D|P = p]$ and $-\mathbb{E}[g(Y)(1 - D)|P = p]$ are non-decreasing in p under Assumptions 2.1, 2.2 and 2.4. The contribution of Theorem 1-(i) is that it shows we do not need to visit every single non-negative measurable function. It is sufficient to restrict our attention to a tractable sub-class of these functions to screen all possible observable violations. This tractable characterization provides a basis for constructing a formal statistical test to verify the validity of the assumptions.

The second part of Theorem 1 is new, and it shows the testable implications in Theorem 1(i) are the most informative way to detect all observable violations of the random assignment, the exclusion restriction, and the monotonicity assumption (without an ex-ante imposed direction). These testable implications cannot be strengthened without making additional assumptions. Various tests or testable implications are used in the literature to screen violations of the judge leniency design assumptions; for instance, Dobbie, Grönqvist, Niknami, Palme, and Priks (2018); Bhuller, Dahl, Løken, and Mogstad (2018); Norris, Pecenco, and Weaver (2021); Frandsen, Lefgren, and Leslie (2023). However, to the best of our knowledge, only Theorem 1 provides sharp testable implications without imposing an a priori direction in the monotonicity assumption.

Tests based on sharp testable implications have empirical virtue. In practice, one may use tests developed from non-sharp testable implications for the sake of traceability. However, as recently discussed in Kédagni, Li, and Mourifié (2020), non-sharp tests can lead to discordant empirical results and misleading interpretations of the estimand of interest. It is possible that for the same data, two different non-sharp tests generate contradictory results as they may use different sets of information from the same observed DGP to screen violations of the model assumptions. Thus, the conclusion may largely depend on which test the empirical researcher implements.

Moreover, after implementing a specification test and obtaining a non-rejection result, one often proceeds and provides a causal interpretation of the estimand. For example, in judge leniency designs, the 2SLS or Local IV (LIV) estimand is interpreted as the LATE or MTE, respectively. However, since a non-sharp test only uses part of the observable information in the data and fails to reject the model when it is misspecified, we must be cautious about interpreting the 2SLS or the Local IV estimand as identifying the LATE/MTE solely based on the result of a non-sharp test. Therefore, using a sharp test must be viewed not only as a theoretical exercise, but also as having important empirical relevance. A sharp test provides the most informative way to detect all observable violations of a given model's assumptions and is more robust to possible misleading interpretations and discordant results.

2.1 Connection to existing tests

2.1.1 Kitagawa (2015), and Mourifié and Wan (2017) testable implications

Inspired by Heckman and Vytlacil (2005, Appendix A), Kitagawa (2015) and Mourifié and Wan (2017) derive a set of sharp testable implications assuming an a priori direction in the monotonicity assumption. When judges' types are binary, i.e. $Z \in \{0, 1\}$, there are only two potential directions, so it is not restrictive to assume the direction of the monotonicity. However, when the cardinality of the judges' types is large (or even infinite when the judges' types are continuous), imposing a specific ex-ante direction is extremely restrictive because the number of possible directions to consider can be rather large (or even infinite). One could implement their test by visiting all the possible directions, but this can be cumbersome or even computationally impossible if Z takes many values.

One significant difference between the testable implication of Kitagawa (2015) and Mourifié and Wan (2017) and ours is we do not assume a prior direction. To illustrate this point, suppopse $\mathcal{Z} = \{z_1, ..., z_K\}$ and suppose we assume one of the $K!$ potential directions as:

$$D_{z_K} \geq D_{z_{K-1}} \geq ... \geq D_{z_1}$$

meaning that type z_K judge is less lenient than type z_{K-1} judge, which, in turn, is less lenient than z_{K-2}, z_{K-3}, $\cdots$, z_1 judge. Assumptions 2.1 to 2.3, plus the above-imposed direction, imply the following testable implications studied in Sun (2023):

$$\begin{aligned}\mathbb{P}(y < Y \leq y', D = 1|Z = z_k) &\leq \mathbb{P}(y < Y \leq y', D = 1|Z = z_{k+1}),\\ -\mathbb{P}(y < Y \leq y', D = 0|Z = z_k) &\leq -\mathbb{P}(y < Y \leq y', D = 0|Z = z_{k+1}),\\ &\text{for all } k \in \{1, ..., K-1\} \text{ and } y, y' \in \mathcal{Y}.\end{aligned}$$

A key point to note is the above implications restrict $F_{Y,D|Z}(y, d|z)$ while the testable implications in Theorem 1(i) instead restrict $F_{Y,D|P}(y, d|p)$. In the first case, the induced direction of inequalities is with respect to the observed judge type Z, while in our case, the inequalities are with respect to the propensity score P, which is obtained without imposing a prior direction. Also, noteworthy is if one takes $y = -\infty$ and $y' = \infty$, the testable implications in Theorem 1(i) no longer have any empirical content. But, the testable implications with an ex-ante monotonicity direction still restrict the propensity scores and the judges' types, i.e., P, and Z, such that

$$\mathbb{P}(D = 1|Z = z_k) \leq \mathbb{P}(D = 1|Z = z_{k+1}), \text{ for all } k \in \{1, ..., K-1\}.$$

Therefore, implementing the testing approaches of Kitagawa (2015) and Mourifié and Wan (2017) may reject the judge leniency design assumptions even if Assumptions 2.1 to 2.3 hold, but just the ex-ante imposed direction of monotonicity is wrong.

2.1.2 Frandsen, Lefgren, and Leslie (2023)'s test

FLL proposes a set of testable implications for Assumptions 2.1 to 2.3. Their testable implication has sound features of not relying on the ex-ante specified direction of monotonicity and assessing all the assumptions jointly. Their testable implication, however, is not sharp and can fail to screen some observable violations of the judge leniency design. To see this, consider any integrable function $g(\cdot) : \mathcal{Y} \to \mathbb{R}$, and let $p \neq p' \in \mathcal{P}$. Under

Assumptions 2.1 to 2.3, we can derive the following equality:

$$\begin{aligned}W(g(Y),p,p') &\equiv \frac{\mathbb{E}[g(Y)|P=p'] - \mathbb{E}[g(Y)|P=p]}{p'-p}\\ &= \mathbb{E}[g(Y_1)-g(Y_0)|p<V\le p']1\{p<p'\} + \mathbb{E}[g(Y_1)-g(Y_0)|p'<V\le p]1\{p<p'\}.\end{aligned}$$

If we denote by L_g and U_g the known lower bound and upper bound of the support of $g(Y)$, the latter equality implies:

$$L_g - U_g \le W(g(Y),p,p') \le U_g - L_g, \tag{2.3}$$

where the inequality in (2.3) is the main testable implication used by FLL (see Theorem 1 and Equation (2) therein) to implement their test. However, under Assumptions 2.1 to 2.3, we should also have:

$$\begin{aligned}W(g(YD),p,p') &= \mathbb{E}[g(Y_1)|p<V\le p']1\{p<p'\} + \mathbb{E}[g(Y_1)|p'<V\le p]1\{p>p'\},\\ W(g(Y(1-D)),p,p') &= -\mathbb{E}[g(Y_0)|p<V\le p']1\{p<p'\} - \mathbb{E}[g(Y_0)|p'<V\le p]1\{p>p'\},\end{aligned}$$

where those two latter equalities lead to the following observable restrictions:

$$L_g \le W(g(YD),p,p') \le U_g, \tag{2.4}$$

$$-U_g \le W(g(Y(1-D)),p,p') \le -L_g, \tag{2.5}$$

One can easily observe that the testable restrictions in (2.4) and (2.5) could be violated, whereas the restriction used by FLL, i.e. inequality (2.3) still holds. These discordant implications confirm the concern about developing a statistical test based on non-sharp restrictions. Hence, implementing FLL's statistical testing procedure based on inequalities (2.4) or (2.5) could provide a different result compared to their test based on inequality (2.3) alone.

Another evident reason why FLL's testing approach cannot exhaust all violations of the judge leniency design is they only focus on $g(Y) = Y$ whereas the inequality in (2.3) should hold for any integrable function g and for any pair $p \neq p' \in \mathcal{P}$. $g(Y) = Y$ is not a sufficient class of functions to screen all violations of the model.

Finally, we note our testable implications in Theorem 1 do not rely on the known support of $g(Y)$, whereas to test inequality (2.3), one needs to know the bounds of the support (U_g, L_g). If the support of $g(Y)$ is unbounded, i.e. $U_g = +\infty$ and $L_g = -\infty$, then the testable implication in (2.3) holds trivially and does not have any power in detecting violations to the identification assumptions.

In the next section, we propose a testing procedure based on the sharp testable implications of Theorem 1. We will show that in large samples, our test is consistent against all the violations of our testable implication and is, therefore, more powerful asymptotically than the existing ones.

3 Testing Procedures

We first present our baseline semi-nonparametric test in Section 3.1 without the presence of control variables X. For this test, we do not make any functional form or distributional assumptions on potential outcomes. For the propensity score, we follow the common practice in the literature to employ a parametric model so that $P(z) = P(z, \theta_0)$ for all $z \in \mathcal{Z}$ and for a finite-dimensional parameter vector $\theta_0 \in \Theta$. Popular choices include the Probit or Logit model with a linear index $z'\theta_0$ (see, for instance, Carneiro, Heckman, and Vytlacil, 2011; Kowalski, 2016, among many others). We also want to emphasize when the instrument variable Z is the judge's identity, we do not need to impose any parametric assumptions on the propensity score. It can be estimated by the sample average of D conditioning on each judge. In this case, our test is indeed nonparametric.[3]

In practice, researchers may observe a set of defendant and case covariates X and assume the randomization and monotonicity hold conditioning on X (see Assumptions 3.1 and 3.2 below). In the presence of covariates, researchers can use the semi-nonparametric test introduced in Section 3.1 when the dimension of covariates is small or the number of support points in $\mathcal{X}$ is not large; please see Remark 3.1 below. In other cases, the semi-nonparametric test may encounter challenges associated with the curse of dimensionality. To address this concern, we introduce an alternative semiparametric test designed to accommodate situations with a large (but fixed) number of covariates in Section 3.2.

[3] When Z is continuous, the rejection result of our semi-nonparametric test can be interpreted as rejecting the joint assumption of the judge leniency design and the parametric form imposed on the propensity score. In our simulation studies, we always keep the propensity score correctly specified. In these studies, therefore, the rejection shows the power of our test to reject false judge leniency assumptions.

3.1 A Semi-nonparametric test

For the convenience of the exposition, we re-state the testable implications as the null hypothesis H_0. That is, for all $p_1 \geq p_2$ with $p_1, p_2 \in \mathcal{P}$ and all $y, y' \in \mathcal{Y}$,

$$\mathbb{P}(y < Y \leq y', D = 1 | P = p_1) \geq \mathbb{P}(y < Y \leq y', D = 1 | P = p_2), \tag{3.1}$$

$$\mathbb{P}(y < Y \leq y', D = 0 | P = p_1) \leq \mathbb{P}(y < Y \leq y', D = 0 | P = p_2). \tag{3.2}$$

The alternative hypothesis H_1 is then inequality (3.1) or (3.2) fails to hold for some (p_1, p_2) and (y, y'). Without loss of generality, we assume the support of Y is $[0, 1]$.[4] Testing inequalities (3.1) and (3.2) involves two features; first, it is a set of inequality restrictions defined on conditional moments where the conditioning variable is possibly continuous. We deal with the first difficulty by employing the method of Hsu, Liu, and Shi (2019) to transform them into an equivalent set of restrictions on unconditional moments. The second feature is the conditioning variable P is not directly observed from the data. For this latter feature, we match our testing procedure with most of the empirical practice and estimate the propensity score parametrically in the first step. We derive the new influence functions and show the first-stage estimation error is properly accounted for.

To be more specific, we define a collection of functions $\{\nu_d(\ell) : \ell \in \mathcal{L}, d = 0, 1\}$ as follows:

$$\begin{aligned} \nu_1(\ell) \equiv \mathbb{E}[D1\{y \leq Y \leq y + r_y\}1\{p_2 \leq P \leq p_2 + r_p\}] \cdot \mathbb{E}[1\{p_1 \leq P \leq p_1 + r_p\}] \\ - \mathbb{E}[D1\{y \leq Y \leq y + r_y\}1\{p_1 \leq P \leq p_1 + r_p\}] \cdot \mathbb{E}[\{p_2 \leq P \leq p_2 + r_p\}], \end{aligned} \tag{3.3}$$

and

$$\begin{aligned} \nu_0(\ell) \equiv \mathbb{E}[(D-1)1\{y \leq Y \leq y + r_y\}1\{p_2 \leq P \leq p_2 + r_p\}] \cdot \mathbb{E}[1\{p_1 \leq P \leq p_1 + r_p\}] \\ - \mathbb{E}[(D-1)1\{y \leq Y \leq y + r_y\}1\{p_1 \leq P \leq p_1 + r_p\}] \cdot \mathbb{E}[1\{p_2 \leq P \leq p_2 + r_p\}], \end{aligned} \tag{3.4}$$

[4] We can always apply a transformation to ensure the support of Y is $[0, 1]$. If Y has a finite support $[a, b]$, we can apply an affine transformation $\tilde{Y} = (Y - a)/(b - a)$. If Y's support is the whole real line, we can apply standard normal CDF after rescaling and recentering: $\tilde{Y} = \Phi\left(\frac{Y - \bar{Y}}{\hat{std}(Y)}\right)$, where $\bar{Y}$ is the sample average and $\hat{std}(Y)$ is the sample standard deviation.

where the index $\ell \in \mathcal{L}$ is defined as

$$\ell = (\ell_y', \ell_p')', \quad \ell_y = (y, r_y)', \quad \ell_p = (p_1, p_2, r_p)', \quad \mathcal{L} = \mathcal{L}_Y \otimes \mathcal{L}_P,$$
$$\mathcal{L}_Y = \left\{(y, r_y) : \ r_y = q_y^{-1}, \ \ q_y \cdot y \in \{0, 1, 2, \cdots, (q_y - 1)\} \text{ for } q_y = 1, 2, \cdots, \right\}.$$
$$\mathcal{L}_P = \left\{(p_1, p_2, r_p) : \ r_p = q_p^{-1}, \ \ q_p \cdot (p_1, p_2) \in \{0, 1, 2, \cdots, (q_p - 1)\}^2, p_1 \geq p_2 \text{ for } q_p = 1, 2, \cdots, \right\}.$$

Then, following the same calculation as in Hsu, Liu, and Shi (2019), we can formulate the null hypothesis in inequalities (3.1) and (3.2) as the following:

$$H_0 : \nu_d(\ell) \leq 0, \quad \text{for all } \ell \in \mathcal{L} \text{ and } d = 0, 1, \tag{3.5}$$

against the alternative hypothesis H_1 that inequality (3.5) fails to hold for some $\ell \in \mathcal{L}$ and for $d = 0$ or $d = 1$. Consequently, testing the original sharp implication in Theorem 1 is equivalent to testing the set of inequalities indexed by $\ell \in \mathcal{L}$, a class of cubes. There is no loss of information for such transformation (see Andrews and Shi, 2013). Under H_0, we expect to see $T \equiv \sum_{d=0,1} \sum_{\ell \in \mathcal{L}} \max\{\nu_d(\ell), 0\}^2 \Omega(\ell) = 0$, where $\Omega(\cdot)$ is a positive weighting function. On the other hand, $T > 0$ under H_1. Our test statistics are based on the appropriately rescaled and standardized sample analog of T.

In the expression of $\nu_d(\ell)$, the propensity score $P(Z, \theta_0)$ is unknown, but can be replaced by its estimate $\hat{P} \equiv P(Z, \hat{\theta})$, where $\hat{\theta}$ is the MLE. Under this parameterization, we also write $\nu_d(\ell)$ as $\nu_d(\ell, \theta_0)$ when it causes no confusion. Algorithm 3.1 below summarizes the semi-nonparametric test's implementation procedure. Please see Appendix A for detailed equations and expressions.

Algorithm 3.1 *This algorithm shows the steps for constructing the test statistics and critical value.*

1. *Specify integers Q_Y and Q_P, and create a coarser version $\mathcal{L}_Q$ of $\mathcal{L}$ set by limiting $q_y = 1, 2, \cdots, Q_Y$ and $q_p = 1, 2, \cdots, Q_P$.*

2. *Estimate the propensity score by $\hat{P}_i = P(Z_i, \hat{\theta})$, where $P(z, \theta)$ parameterized up to θ and $\hat{\theta}$ is the maximum likelihood estimator.*

3. *For each $\ell \in \mathcal{L}_Q$, construct estimates $\hat{\nu}_1(\ell)$ and $\hat{\nu}_0(\ell)$ as sample analog of Equations (3.3) and (3.4).*

4. *Choose a constant $\epsilon > 0$. For each $\ell \in \mathcal{L}_Q$, construct $\hat{\sigma}^2_{d,\epsilon}(\ell) = \max\{\hat{\sigma}^2_d(\ell), \epsilon\}$, where $\hat{\sigma}^2_d(\ell)$ is a consistent estimator for the asymptotic variance of $\sqrt{n}(\widehat{\nu}_d(\ell, \widehat{\theta}) - \nu_d(\ell, \theta_0))$ and is defined below in Equation (3.8).*

5. *Choose the weighting function Ω over $\mathcal{L}$ such that $\Omega(\ell) > 0$ for all $\ell \in \mathcal{L}$ and $\sum_{\ell \in \mathcal{L}} \Omega(\ell) < \infty$. Calculate the test statistics as*

$$\widehat{T}_n = \sum_{d=0,1} \sum_{\ell \in \mathcal{L}_Q} \max\left\{\sqrt{n}\frac{\hat{\nu}_d(\ell)}{\hat{\sigma}_{d,\epsilon}(\ell)}, 0\right\}^2 \Omega(\ell).^{5} \tag{3.6}$$

6. *Let a_n and B_n be positive deterministic sequences.*[6] *Calculate the generalized moment selection (GMS) terms as*

$$\hat{\psi}_d(\ell) = -B_n \cdot 1\left\{\frac{\sqrt{n}\hat{\nu}_d(\ell)}{\hat{\sigma}_{d,\epsilon}(\ell)} < -a_n\right\}.$$

7. *Choose a positive integer B (as the number of bootstrap iterations), and for each $b = 1, 2, \cdots, B$,*

   (a) *Draw $W^b_1, W^b_2, \cdots, W^b_n$ as a sequence of independent random variables with both mean and variance equal to one and are independent of the original sample.*

   (b) *Estimate propensity score for each bootstrap iteration $\hat{P}^b_i = P(Z_i, \hat{\theta}^b)$, where $\hat{\theta}^b$ is the maximum likelihood estimator.*

   (c) *Obtain $\hat{\nu}^b_d(\ell)$, $d = 0, 1$, for each bootstrap iteration using Equations (A.2) and (A.3).*

   (d) *Calculate the quantity*

$$\widehat{T}^b = \sum_{d \in \{0,1\}, \ell \in \mathcal{L}_Q} \max\left\{\frac{\hat{\Phi}^b_d(\ell)}{\hat{\sigma}_{d,\epsilon}(\ell)} + \hat{\psi}_d(\ell)\right\}^2 \Omega(\ell),$$

   where

$$\Phi^b_d(\ell) = \sqrt{n}\left(\hat{\nu}^b_d(\ell) - \hat{\nu}_d(\ell)\right). \tag{3.7}$$

[5] To be specific, for q_y and q_p, we suggest to set $\Omega(\ell) = q_y^{-3} \cdot \frac{q_p^{-2}}{q_p(q_p-1)}$.

[6] See Andrews and Shi (2013) for the rate condition of a_n and B_n and they suggest to set $a_n = \sqrt{0.3 \ln n}$ and $B_n = \sqrt{0.4 \ln n / \ln \ln n}$. Here, we propose $a_n = 0.15 \ln n$ and $B_n = 0.85 \ln n / \ln \ln n$, as in Hsu, Liu, and Shi (2019).

8. *Estimate $\hat{\sigma}_d(\ell)$ by*

$$\hat{\sigma}_d^2(\ell) = \frac{n}{B}\sum_{b=1}^{B}\left(\hat{\nu}_d^b(\ell) - \overline{\hat{\nu}_d^b}(\ell)\right)^2, \text{ where } \overline{\hat{\nu}}_d^b(\ell) = \frac{1}{B}\sum_{b=1}^{B}\hat{\nu}_d^b(\ell). \tag{3.8}$$

9. *Let $\hat{c} = \hat{q}(1-\alpha+\eta)+\eta$, where $\hat{q}(\tau)$ is the τ-th empirical quantile of $\left\{\widehat{T}^b\right\}_{b=1}^{B}$ and η is a small positive constant, e.g. $\eta = 10^{-6}$.*[7]

10. *Define the test to be $\phi_n = 1\{\widehat{T} \geq \hat{c}\}$. That is, we reject the null hypothesis if $\widehat{T} \geq \hat{c}$.*

Theorem 2 shows that the test ϕ_n has its size controlled asymptotically and is consistent. The proof for Theorem 2 is collected in Section C of the online supplementary material. We also list all the technical conditions in that section for exposition purposes.

Theorem 2 *Suppose Assumptions C.1 to C.4 in Appendix D are satisfied. Let $\alpha \in (0, 1/2)$ be the pre-chosen significance level.*

(i) Under the H_0 in characterized by inequalities (3.5), we have

$$\limsup_{n\to\infty} \mathbb{P}(\phi_n = 1|H_0) \leq \alpha. \tag{3.9}$$

(ii) Under H_1,

$$\limsup_{n\to\infty} \mathbb{P}(\phi_n = 1|H_0) = 1. \tag{3.10}$$

3.2 A semiparametric test with covariates dimension reduction

In this section, we introduce a semiparametric test in the presence of covariates X. We begin by introducing the following assumptions.

Assumption 3.1 (Conditional Random Assignment of Judges) *$Z \perp (Y_0(z), Y_1(z), D(z); z \in \mathcal{Z})|X = x$ for all $x \in \mathcal{X}$*

Assumption 3.2 (Single Threshold-Crossing with Covaraites: STC) *The judge treatment assignment mechanism is governed by the following threshold crossing model $D =$*

[7] η is the so infinitesimal constant which is introduced mainly for the sake of proof; see for instance Andrews and Shi (2013). Our simulation exercises set it to 10^{-6}.

$1\{\nu(Z, X) \geq U\}$ *for some measurable and non-trivial function* ν, *where the distribution of* U *is absolutely continuous.*

When Assumptions 2.2, 3.1 and 3.2 hold, the testable implications can be written as follows. For all $x \in \mathcal{X}$, $p_1, p_2 \in \mathcal{P}$ and $p_1 \leq p_2$, and all $y, y' \in \mathcal{Y}$

$$\mathbb{P}(y < Y \leq y', D = 1 | P = p_1, X = x) \geq \mathbb{P}(y < Y \leq y', D = 1 | P = p_2, X = x), \quad (3.11)$$

$$\mathbb{P}(y < Y \leq y', D = 0 | P = p_1, X = x) \leq \mathbb{P}(y < Y \leq y', D = 0 | P = p_2, X = x). \quad (3.12)$$

Remark 3.1 *If* X *is discrete and* $\mathcal{X}$ *only contains a relatively small number of values, we can implement the procedure in Section 3.1 for each subsample defined by those support points, and control the familywise error rate by certain multiple testing procedure, e.g. Holm (1979). If* X *contains a small number of continuous variables, we can also follow the same procedure as in Section 3.1 but adding cubes for* X*. For example, it implies* $\nu_1(\ell, x) \leq 0$ *for all* (ℓ, x) *where*

$$\begin{aligned} \nu_1(\ell, x) \equiv {} & \mathbb{E}[D1\{y \leq Y \leq y + r_y\}1\{x \leq X \leq x + r_x\}1\{p_2 \leq P \leq p_2 + r_p\}] \\ & \times \mathbb{E}[1\{x \leq X \leq x + r_x\}1\{p_1 \leq P \leq p_1 + r_p\}] - \mathbb{E}[1\{x \leq X \leq x + r_x\}\{p_2 \leq P \leq p_2 + r_p\}] \\ & \times \mathbb{E}[D1\{y \leq Y \leq y + r_y\}1\{x \leq X \leq x + r_x\}1\{p_1 \leq P \leq p_1 + r_p\}], \end{aligned}$$

and

$$\begin{aligned} \nu_0(\ell) \equiv {} & \mathbb{E}[(D-1)1\{y \leq Y \leq y + r_y\}1\{x \leq X \leq x + r_x\}1\{p_2 \leq P \leq p_2 + r_p\}] \\ & \times \mathbb{E}[1\{x \leq X \leq x + r_x\}1\{p_1 \leq P \leq p_1 + r_p\}] - \mathbb{E}[1\{x \leq X \leq x + r_x\}1\{p_2 \leq P \leq p_2 + r_p\}] \\ & \times E[(D-1)1\{y \leq Y \leq y + r_y\}1\{x \leq X \leq x + r_x\}1\{p_1 \leq P \leq p_1 + r_p\}], \end{aligned}$$

and r_x *is similarly defined as* r_p *and* r_y*. The implementation follows analously from Algorithm 3.1.*

When the dimension of X is high, an alternative approach is to include the covariates parametrically, as in Carr and Kitagawa (2021, Assumptions A.4 and A.5), which we state below:

Assumption 3.3 *(i) For $d = 0, 1$, then potential outcomes take the form of $Y_d = \alpha_d + X'\beta_d + U_d$, where (α_d, β_d) are constants, and (ii) the residual terms (U_0, U_1) satisfy $(U_0, U_1, V) \perp (X, Z)$.*

Carr and Kitagawa (2021, Proposition 2) show if Assumption 3.1 is strengthened to Assumption 3.3, then the testable implications in (3.11) and (3.12) can be characterized as

$$\mathbb{P}(y < \tilde{Y} \leq y', D = 1 | P = p_1) \geq \mathbb{P}(y < \tilde{Y} \leq y', D = 1 | P = p_2), \tag{3.13}$$

$$\mathbb{P}(y < \tilde{Y} \leq y', D = 0 | P = p_1) \leq \mathbb{P}(y < \tilde{Y} \leq y', D = 0 | P = p_2), \tag{3.14}$$

for $y, y' \in \mathcal{Y}$, and

$$\tilde{Y} = D(U_1 + \alpha_1) + (1 - D)(U_0 + \alpha_0) = D(Y_1 - X'\beta_1) + (1 - D)(Y_0 - X'\beta_0) = Y - X'(D\beta_1 + (1 - D)\beta_0).$$

The advantage of using (3.13) and (3.14) is both inequalities are only conditional on the scalar-valued propensity score. The effect of covariates has been filtered out by constructing a new outcome variable $\tilde{Y}$. Under the null hypothesis of the model being correctly specified, parameters β_0 and β_1 can be estimated by partial linear regression of Y on X and propensity score P separately for the sample of $D = 1$ and $D = 0$; see for instance Carneiro and Lee (2009); Carneiro, Heckman, and Vytlacil (2010); Kowalski (2016). Specifically, for the potential outcome Y_d, it can be shown that

$$\mathbb{E}[Y | X = x, P = p, D = d] = x'\beta_d + K_d(p),$$

where $K_d(p) = \mathbb{E}[\alpha_d + U_d | X = x, D = d, P = p]$ only depends on p under Assumption 3.3-(ii). The following algorithm summarizes the steps for implementation.

Algorithm 3.2 *1. The procedure starts with estimated propensity score $\hat{P}_i = P(Z_i, X_i, \hat{\theta})$ using Equation (E.1).*

2. Choosing the subsample with $D = d$, and within this subsample,

(a) Estimate $\mathbb{E}[Y|P]$ nonparametrically,[8] and calculate the residual $e_i^P \equiv Y_i - \hat{\mathbb{E}}[Y_i | \hat{P}_i]$.

[8] One can consider local polynomial estimation as in Carneiro and Lee (2009) or do global estimation

(b) *Estimate* $\mathbb{E}[X|P]$ *nonparametrically, and calculate the residual* $e_i^X \equiv X_i - \hat{\mathbb{E}}[X_i|\hat{P}_i]$.

(c) *Regress* e_i^P *on* e_i^X *and obtain the OLS estimates, denoted by* $\hat{\beta}_d$.

3. *Once* $\hat{\beta}_1$ *and* $\hat{\beta}_0$ *are obtained, one can construct estiamtes for* $\widetilde{Y}_i = Y_i - X_i'(D_i\hat{\beta}_1 + (1 - D_i)\hat{\beta}_0)$

4. *Follow the rest of steps in Algorithm 3.1 with* Y *being replaced by* $\widetilde{Y}$.

4 Simulation and Empirical Application

4.1 Simulation

In this subsection, we provide two sets of simulations to assess the size and power properties of our sharp test under various DGPs in finite samples. Throughout this section, we ran 1000 replications for each simulation design, and the bootstrap sample size is chosen to be $B = 800$. We set $a_n = 0.15 \ln n$ and $B_n = 0.85 \ln n / \ln \ln n$, as in Hsu, Liu, and Shi (2019). We choose $Q_P = 5$ and $Q_Y = 5$ (for continuous Y) or $Q_Y = 2$ (for binary Y). We set the infinitesimal constant $\eta = 10^{-6}$ and the constant $\epsilon = 10^{-6}$ (see the definition of $\hat{\sigma}^2_{d,\epsilon}(\ell)$ in Algorithm 3.1-4). The simulation results are not sensitive to these constants.

4.1.1 Binary outcome

The first set of simulations is based on a DGP introduced in FLL (online appendix, page 22). In this set of simulations, we mimic the random assignment of n defendants to a pool of J judges, ensuring an equitable distribution of $\frac{n}{J}$ defendants to each judge. As in FLL, the severity probability of each judge j is set as follows:

$$p_j = p_a + \frac{j-1}{J-1}(1 - p_a - p_n)$$

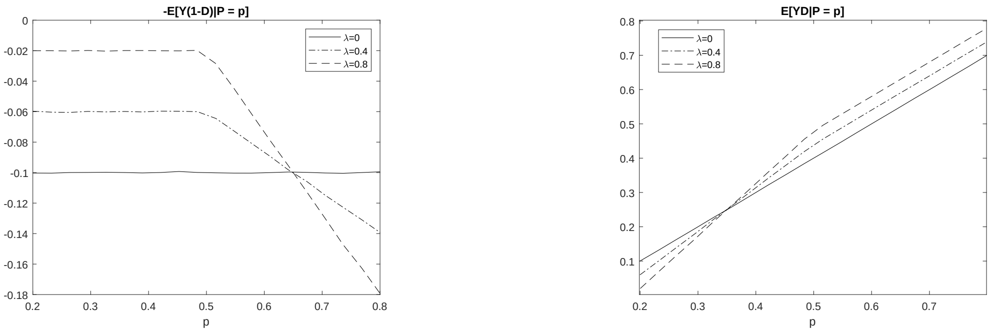
Here, p_a and p_n stand for the fraction of always and never treated defendants, respectively. FLL consider a binary outcome model where the outcome $Y \in \{0, 1\}$ satisfies the following

as in Kowalski (2016). Since we do not need to estimate the derivative $K'(p)$ in our paper, we use global polynomial regression in Kowalski (2016) for our simulation and empirical applications.

condition:

$$\mathbb{E}\left[Y \mid p_j\right] = \frac{1-(1-\lambda)(p_n+p_a)}{1-(p_n+p_a)}p_j - \frac{\lambda}{1-(p_n+p_a)}p_a.$$

The parameter λ dictates the extent of deviation from the exclusion restriction assumption. When $\lambda = 0$, there is no violation of the judge leniency design assumptions. Consequently, for $\lambda = 0$, the simulations aim at assessing the size property of the two different tests. On the other hand, $\lambda > 0$ signifies a departure from the exclusion restriction assumption, with higher (absolute) values indicating a more pronounced deviation. Like in the original paper, we adopt the parametrization for the fraction of always and never treated $p_n = p_a = 0.2$. Meanwhile, we vary the value of λ within the range of 0 to 1.



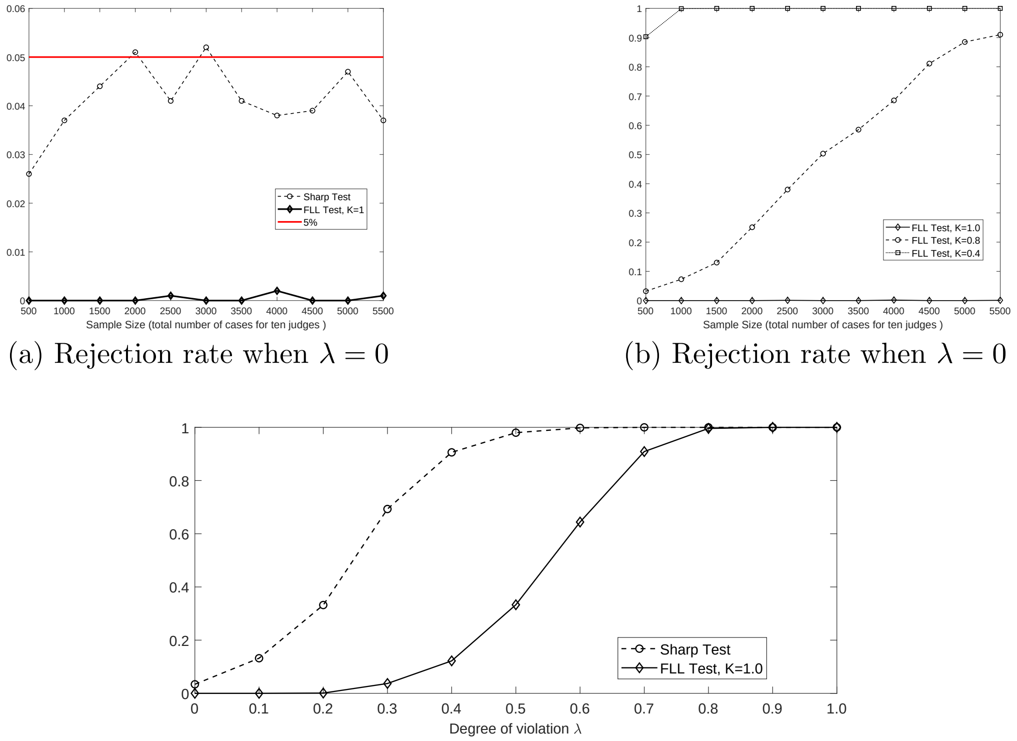
(a) $-\mathbb{E}[Y(1-D)|P=p]$ (b) $\mathbb{E}[YD|P=p]$

Figure 1: Testable restrictions by degree of violations of exclusion restriction

Figure 1 visually illustrates our testable implications of the judge leniency design for the specific function $g(Y) = 1\{0 < Y \leq 1\} = Y$ (because Y is binary). The left and right panels of the figure, respectively, depict $\mathbb{E}[-Y(1-D)|P=p]$ and $\mathbb{E}[YD|P=p]$. These population quantities are approximated by a large number of defendants (1 million) for each judge. Intuitively, it is expected that $\mathbb{E}[YD|P=p]$ and $\mathbb{E}[-Y(1-D)|P=p]$ should be non-decreasing when the judge leniency design holds. When the exclusion restriction holds, as shown in both figures with $\lambda = 0$, $\mathbb{E}[YD|P=p]$ and $\mathbb{E}[-Y(1-D)|P=p]$ behave as expected. However, for a violation of the exclusion restriction ($\lambda = 0.4$ or $\lambda = 0.8$), despite that $\mathbb{E}[YD|P=p]$ remains to be increasing, the other function $\mathbb{E}[-Y(1-D)|P=p]$ decreases for higher values of the propensity score. This discrepancy starkly contrasts with the implications of the judge leniency design assumptions.

In Figure 2-(a), we report the size property for our sharp test and FFL's test at 5%

significance level (when $\lambda = 0$). The simulation designs involve twenty judges and varying sample sizes, ranging from 500 defendants (equivalent to 50 defendants per judge) to 5500 defendants (equivalent to 550 defendants per judge). The plot reveals both tests control size well in the aforementioned DGP. Specifically, it is evident from the graph the rejection rate of our sharp test is controlled by and close to the nominal level 5%. Conversely, the nonparametric test proposed in FLL consistently yields rejection rates close to zero when setting the tuning parameter $K = 1$.[9]



(a) Rejection rate when $\lambda = 0$

(b) Rejection rate when $\lambda = 0$

(c) Rejection rate with varying degree of violation ($n = 1000$)

Figure 2: Rejection rates in FLL's DGP

FLL discuss how one can improve the power of their testing methodology by considering more stringent upper bounds on the largest possible treatment effects (i.e. using a smaller value of K). For instance, in their empirical application of a binary outcome model–where the maximum treatment effect is set at 1–they advocate exploring smaller permissible maximum treatment effect values. However, if K is set to be too small, then FLL's test can have server size distortion. Indeed, Figure 2-(b) graphically represents this situation by plotting the rejection rate associated with FLL's nonparametric test under

[9] Recall the outcome variable is binary; hence, the largest possible absolute value for the treatment effect is 1.

two additional cases: when the maximum allowable treatment effect K is set at 0.8 and 0.4, respectively. The striking observation is the conclusions drawn from these scenarios can be misleading, as they suggest an excessive over-rejection of the assumptions even when those assumptions are indeed satisfied. For example, if one sets $K = 0.4$, then the rejection rate is always 100% whenever the sample size is greater or equal to 1000. As a matter of fact, the rejection we observe from Figure 2-(b) reflects that the ad-hoc imposed magnitude of the treatment effect is not correct, but the underlying exclusion restriction holds. Our test is immune to this problem since it does not require pre-specifying the magnitude order of the unknown treatment effect.

To assess and compare the power property of the two nonparametric tests, Figure 2-(c) plots the rejection rate as a function of λ for 10 judges and 1000 defendants (100 defendants per judge). The solid line is the rejection rate of the FLL test, which is nearly the same as what is plotted in FLL (Appendix, Figure 10). The rejection rate achieved by our sharp test consistently surpasses that of the FLL test across the entire spectrum of exclusion restriction violations, as indicated by varying degrees of λ. As shown, the power improvement can be substantial.

4.1.2 Continuous outcome

The second set of simulations aims to show the performance of our test in detecting violations of the judge leniency design when the outcome is continuous and unbounded. Let $(U_0, U_1, U, Z^*) \sim N(\boldsymbol{\mu}, \Sigma)$, where $\boldsymbol{\mu} = (\mu_0, \mu_1, \mu_U, \mu_Z)'$ is a vector of means, and Σ is a covariance matrix. For generic random variables A and B, let σ_A^2 be the variance of A and $\rho_{A,B}$ be the correlation coefficient between A and B. In this design, we set $\sigma_A = 1$ for all $A \in \{U_1, U_0, U, Z^*\}$. We let $\rho_{U_0,U} = -0.5$, $\rho_{U_1,U} = 0.5$, $\rho_{U,Z} = 0$, $\rho_{U_1,U_0} = 0$, $\rho_{U_1,Z} = \delta_1$, and $\rho_{U_0,Z} = \delta_1$. To create discrete judges or IV Z, we set

$$Z = F_{Z^*}^{-1}\left(\min\left\{\ell = 1, 2, \cdots, L-1 : \left|F_{Z^*}(Z^*) - \frac{\ell}{L}\right|\right\}\right),$$

that is, we divide the support of Z^* by L equal-probability intervals and concentrate the mass over each interval to its nearest cutoff points. Let the potential outcomes and treatment assignment be

$$\begin{aligned} D =& 1\{\nu(X,Z) > U\} \times 1\{\delta_2 = 0\} \\ &+ [1\{\nu(X,Z) > U\}1\{U \geq U_0\} + 1\{1-\nu(X,Z) > U\}1\{U < U_0\}] \times 1\{\delta_2 \neq 0\}, \end{aligned}$$

and

$$Y_d(z) = \alpha_d + X\beta_d + \delta_3 z + U_d, \quad Y_d = \sum_{z\in\mathcal{Z}} Y_d(z)1\{Z = z\}.$$

where $X \sim N(0,1)$ is independent of all the other random variables. We let $\nu(x,z) = z$ and set $\alpha_0 = 0$, and $\alpha_1 = 1$.The δ parameters, however, are set to be different values to capture different violations of the judge leniency design. More specifically,

1. when $\delta_1 = \delta_2 = \delta_3 = 0$, the assumptions of the judge leniency design hold;

2. $\delta_1 \neq 0$ denotes violation of the independence assumption;

3. $\delta_2 \neq 0$ denotes violation of the monotonicity assumption; In this case, the selection equation becomes

   $$D = 1\{Z > U\}1\{U \geq U_0\} + 1\{1-Z > U\}1\{U < U_0\},$$

   which indicates that there are two groups of judges, with each group having distinct skills (or preferences) in assigning treatment. This is in clear violation of the monotonicity assumption, which requires all judges to have the same skill (Chan, Gentzkow, and Yu, 2022).

4. $\delta_3 \neq 0$ denotes violation of the exclusion restriction.

Figure 3 plots $\mathbb{E}[g(Y)D|P = p]$ as a function of p when $g(Y) = 1\{Y \geq 0.5\}$ and 20 judges for a simple illustration. The graphs were simulated with a large sample size (over three million) and approximated the population quantity. The function is non-decreasing when all assumptions are met, as shown in the upper-left panel. In contrast, $\mathbb{E}[g(Y)D|P = p]$ deviates from the expected pattern when the judge leniency design assumptions are violated in different ways.

Figure 4, on the other hand, plots the testable implication used in FLL. The left side panels plot $\mathbb{E}[Y|P = p]$ for each of the $p \in \{p_1, p_2, \cdots, p_{20}\}$ (sorted in increasing order)

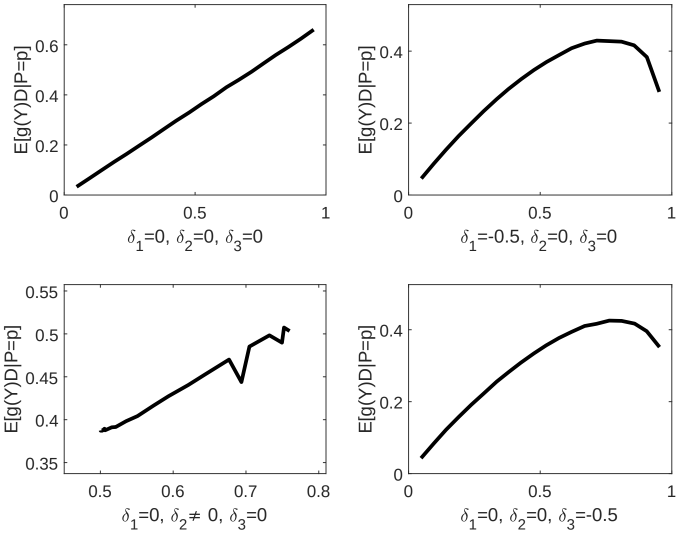


Figure 3: Sharp Testable Restrictions for Different DGPs

for each of the four designs. The right panels plot the "numerical derivative" of the form $\frac{\mathbb{E}[Y|P=p_j]-\mathbb{E}[Y|P=p_{j-1}]}{p_j-p_{j-1}}$ against $\{p_2, \cdots, p_{20}\}$. The FLL testable implications require that the curves in the right-hand side panels be bounded between $[-K, K]$, where K again is the difference between the upper and lower bounds of the support. Note that in this example, the outcomes have unbounded support and, therefore, $K = +\infty$. If we choose K as a large number, then it is apparent all four designs satisfy FLL's testable implication. Hence, we expect no rejection for designs 2-4 albeit they violate the identifying assumptions unless K is set to be relatively small.

We proceed by implementing our sharp test and FLL's nonparametric test. This comparison is conducted across various parameter values and sample sizes. Specifically, we consider a size design (**Size** $\delta_1 = \delta_2 = \delta_3 = 0$), violation of independence (**Power1** $\delta_1 = -0.5, \delta_2 = \delta_3 = 0$), violation of monotonicity (**Power2** $\delta_2 \neq 0, \delta_1 = \delta_3 = 0$), and violation of exclusion (**Power3** $\delta_3 = -0.5, \delta_1 = \delta_2 = 0$). For each violation, we consider situations with covariates ($\beta_1 = \beta_0 = 1$) or without covariates ($\beta_1 = \beta_0 = 0$). To implement FLL's test, we set K to be the difference between sample maximum (y_{max}) and minimum (y_{min}): $\Delta_y \equiv y_{max} - y_{min}$. We also consider $K = \frac{\Delta_y}{8}$ and $K = \frac{\Delta_y}{16}$. The results are summarized in Table 1.

Regarding the size property, all tests control the size except the FLL test when K is set to be very small. Our test and the FLL test with $K = \Delta_y$ and $K = \frac{\Delta_y}{8}$ are conservative. When one sets $K = \frac{\Delta_y}{16}$, the rejection probability of FLL's test increases quickly even when all the assumptions are satisfied (the first design). This is unsurprising

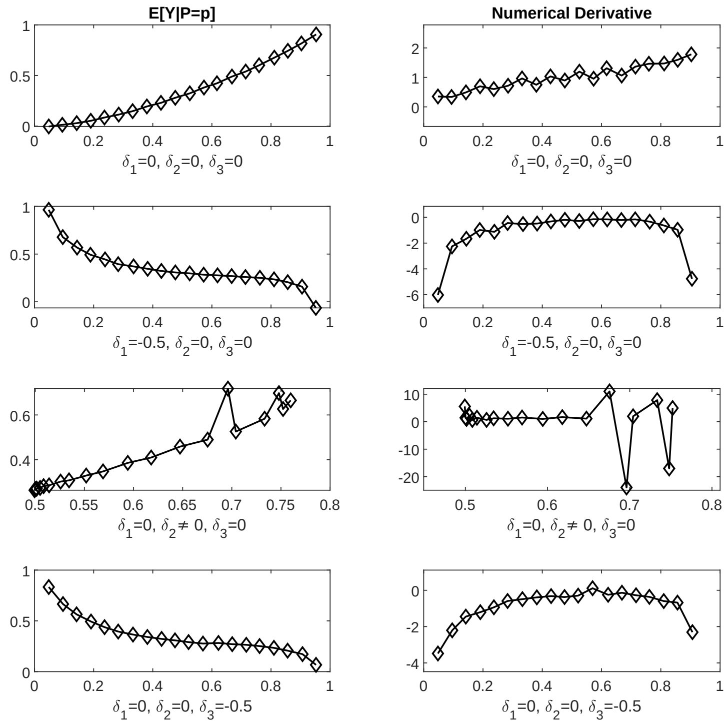


Figure 4: FLL Testable Restrictions for Different DGPs

because a very small K essentially introduced another severe misspecification to the model. However, when examining the power property of the three tests, we see clearly that our test outperforms FLL's tests by a large margin. The proposed sharp test has enough power to detect the violation of any of the three assumptions (independence, exclusion and monotonicity). In particular, the rejection rates for our sharp test quickly increase with sample size, surpassing 90% for all cases when the sample size reaches 2000 (or 100 cases per judge). Note that in this simulation, the parametric form of the propensity score is correctly specified (except for Power2 when monotonicity is violated); hence, the high power of our test is not because of misspecification of $P(z, \theta_0)$. In contrast, FLL's test has low power performance unless we set K as a small value, which, on the other hand, induces size distortion.

Table 2 further examines how the rejection frequency varies as the "magnitude of violation varies" for independence and exclusion. For this exercise, we focus on sample size $n = 1000$ (50 cases per judge). Not surprisingly, when the magnitude of the violation is small, all tests have low power. However, as the degree of violation increases, the

Table 1: Rejection Rate under Different Types of DGPs

	$\delta_1 = \delta_2 = \delta_3 = 0$ (**Size**)			$\delta_1 = -0.5, \delta_2 = \delta_3 = 0$ (**Power1**)		
Without Covariates	$n = 500$	$n = 1000$	$n = 2000$	$n = 500$	$n = 1000$	$n = 2000$
Sharp Test	0.000	0.000	0.000	0.436	0.848	0.995
FLL-nonp, $K = \Delta_y$	0.000	0.000	0.000	0.000	0.000	0.000
FLL-nonp, $K = \frac{\Delta_y}{8}$	0.007	0.001	0.018	0.015	0.054	0.129
FLL-nonp, $K = \frac{\Delta_y}{16}$	0.064	0.284	0.719	0.101	0.376	0.839

	$\delta_2 \neq 0, \delta_1 = \delta_3 = 0$ (**Power2**)			$\delta_3 = -0.5, \delta_1 = \delta_2 = 0$ (**Power3**)		
Without Covariates	$n = 500$	$n = 1000$	$n = 2000$	$n = 500$	$n = 1000$	$n = 2000$
Sharp Test	0.374	0.734	0.942	0.183	0.503	0.902
FLL-nonp, $K = \Delta_y$	0.000	0.000	0.000	0.000	0.000	0.000
FLL-nonp, $K = \frac{\Delta_y}{8}$	0.015	0.037	0.079	0.005	0.004	0.008
FLL-nonp, $K = \frac{\Delta_y}{16}$	0.065	0.104	0.322	0.019	0.049	0.107

	$\delta_1 = \delta_2 = \delta_3 = 0$ (**Size**)			$\delta_1 = -0.5, \delta_2 = \delta_3 = 0$ (**Power1**)		
With Covariates	$n = 500$	$n = 1000$	$n = 2000$	$n = 500$	$n = 1000$	$n = 2000$
Sharp Test	0.000	0.000	0.000	0.424	0.821	0.993
FLL-nonp, $K = \Delta_y$	0.000	0.000	0.000	0.000	0.000	0.000
FLL-nonp, $K = \frac{\Delta_y}{8}$	0.003	0.000	0.000	0.029	0.018	0.041
FLL-nonp, $K = \frac{\Delta_y}{16}$	0.069	0.113	0.293	0.084	0.173	0.456

	$\delta_2 \neq 0, \delta_1 = \delta_3 = 0$ (**Power2**)			$\delta_3 = -0.5, \delta_1 = \delta_2 = 0$ (**Power3**)		
With Covariates	$n = 500$	$n = 1000$	$n = 2000$	$n = 500$	$n = 1000$	$n = 2000$
Sharp Test	0.345	0.714	0.936	0.167	0.488	0.902
FLL-nonp, $K = \Delta_y$	0.000	0.000	0.000	0.000	0.000	0.000
FLL-nonp, $K = \frac{\Delta_y}{8}$	0.006	0.013	0.022	0.004	0.002	0.001
FLL-nonp, $K = \frac{\Delta_y}{16}$	0.050	0.075	0.225	0.018	0.017	0.042

power of our sharp test rises quickly, even quicker than the FLL's nonparametric test with $K = \frac{\Delta_y}{16}$. On the other hand, when $K = \Delta_y$, FLL's nonparametric test does not reject even if the degree of violation is substantial. Again, this table demonstrates that sharp testable implications are desirable in practice.

Table 2: Rejection Rate under Different Levels of Violations (No Covariates)

$\delta_2 = \delta_3 = 0$, $n = 1000$	$\delta_1 = -0.1$	$\delta_1 = -0.3$	$\delta_1 = -0.5$	$\delta_1 = -0.7$
Sharp Test	0.001	0.085	0.825	1.000
FLL-nonp, $K = \Delta_y$	0.000	0.000	0.000	0.000
FLL-nonp, $K = \frac{\Delta_y}{8}$	0.001	0.004	0.054	0.911
FLL-nonp, $K = \frac{\Delta_y}{16}$	0.026	0.006	0.397	0.917

$\delta_1 = \delta_2 = 0$, $n = 1000$	$\delta_3 = -0.1$	$\delta_3 = -0.3$	$\delta_3 = -0.5$	$\delta_3 = -0.7$
Sharp Test	0.000	0.069	0.471	0.931
FLL-nonp, $K = \Delta_y$	0.000	0.000	0.000	0.000
FLL-nonp, $K = \frac{\Delta_y}{8}$	0.000	0.000	0.005	0.114
FLL-nonp, $K = \frac{\Delta_y}{16}$	0.027	0.002	0.032	0.798

4.2 Empirical illustration

In this subsection, we employ our test to assess the validity of the judge leniency designs using data from Stevenson (2018); see also Cunningham (2021), who studies the impact of pretrial detention on conviction. Using Philadelphia court records and leveraging the varying leniency of bail magistrates as an instrumental variable, the author discovers that pretrial detention leads to a 13% increase in the likelihood of conviction.

In the Philadelphia court system, following an arrest, individuals are taken to one of seven city police stations for a videoconference interview by Pretrial Services, assessing risk factors and financial details for public defence eligibility. Utilizing this information, Pretrial Services assigns arrestees to a bail recommendation grid. Bail hearings, conducted by magistrates every four hours over videoconference, involve a brief process where charges are explained, next court appearances are specified, eligibility for a court-appointed defence attorney is determined, and bail amounts are set based on arrest details, interviews, criminal history, guidelines, and input from representatives. Magistrates hold broad authority to assign bail, which can fall into categories such as release without payment, cash bail with a 10% deposit, or no bail at all.

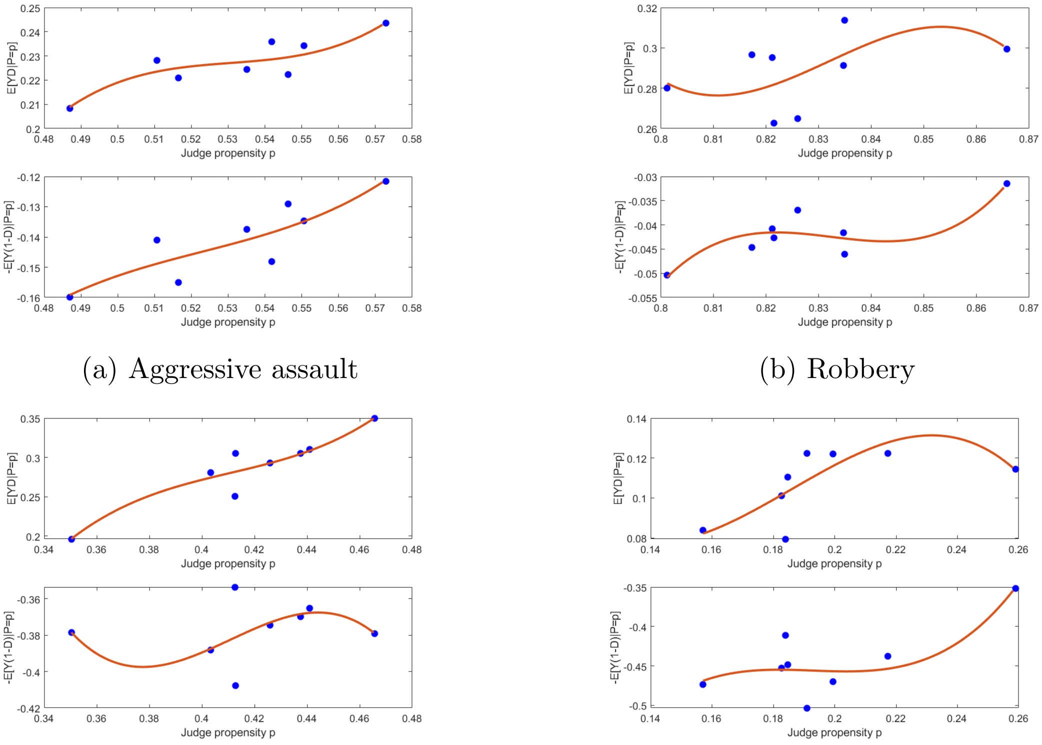
Stevenson (2018)'s research design leverages the varying magistrate tendencies to assign affordable bail as an instrument to study detention's impact on cases' outcomes. To answer the research questions, the author uses data from the court records of the Pennsylvania Unified Judicial System, obtained through web-scraping of public records in PDF format, which are then transformed for statistical analysis. The dataset encompasses arrests in Philadelphia, where charges were filed between September 13, 2006, and February 18, 2013. The final dataset includes 331,971 cases and eight *randomly* assigned judges, with each observation pertaining to a specific criminal case.

In what follows, we focus on the aggregate dataset (all criminal cases together) and four primary categories of criminal cases in the data: aggressive assault, robbery, drug sale, and drug possession. These four criminal cases we consider in isolation constitute 43% of the total cases. In Figure 5, we present a scatter plot with a fitted polynomial to illustrate whether the anticipated implications of the judge leniency design framework are satisfied for the considered categories of criminal cases. The graphs indicate $\mathbb{E}[YD|P=p]$ and $\mathbb{E}[-Y(1-D)|P=p]$ are most likely to be non-decreasing for the aggressive assault case.[10] The non-decreasing shape of the functions is unclear for the other types of criminal categories. Although this graphical representation does not constitute a formal test, it offers an intuitive insight. Specifically, it suggests the assumptions are the least likely to be violated for the aggressive assault case, while the drug possession case shows the highest likelihood of violations of the assumptions of the judge leniency design.

We observe a relatively large set of covariates, including fixed effects for year, month, and day of the week. We, consequently, implement the semi-parametric version of our test. For comparison, we also implement FLL's nonparametric and semi-parametric tests. The results of the three tests are presented in Table 3 for both the aggregate dataset and separately for each of the four crime categories aforementioned. The nonparametric test introduced by FLL indicates the validity of judge leniency design cannot be rejected either conditioning on each crime category or the aggregate data set at 10% level, despite that the shape of $\mathbb{E}[YD|P=p]$ and $\mathbb{E}[-Y(1-D)|P=p]$ for the drug possession type suggests the opposite. In contrast, our novel test yields results that align with expectations. For instance, the sharp test does not indicate a rejection of the validity of the judge leniency design assumptions for the aggressive assault. However, for all three other types of offences, our test rejects the validity of the judge leniency design. Meanwhile, FLL's semi-parametric test rejects the category of aggregate assault.[11] These results suggest that

[10] Note all the outcome variables are binary. Therefore, the close interval we use for the Theorem 1 is $1\{0 < Y \leq 1\}$, which equals to Y.

[11] For FLL's semi-parametric test, we fit the regression function $\mathbb{E}[Y|P=p]$ by B-spline with three knots. The results for other numbers of knots are reported in the appendix. The reported p-value is the "combined p-value" of the fit component and slope component of the test, and we can see from Table 4



(a) Aggressive assault

(b) Robbery

(c) Drug sell

(d) Drug possession

Figure 5: Testable restrictions by case types

using the Wald estimand or the MTE approach for those cases will lead to inconsistent estimates of the causal effects of interest.

Finally, we see no evidence to refute the assumptions underpinning the judge leniency design when applying our sharp test to the aggregate data set. This outcome may be influenced by the notably high proportion of aggressive assault cases within the dataset compared to other categories. Our result also ascertains that the exclusion restriction or monotonicity can hold for some crime categories but not others, suggesting controlling the crime type is important in practice.

5 Partial Exclusion and Monotonicity

The rejection of the sharp test means the judge's leniency design assumptions are too stringent for the data. In these cases, it may be desirable to relax some of these assump-

in the online appendix that the rejection is mostly generated by the fit component.

Table 3: Testing Judge Leniency Design: p-values

	Sharp Test	FLL-Nonp	FLL-Semip
All	0.821	0.056	0.114
Aggressive assault	0.913	0.996	0.015
Robbery	0.033	1.000	0.109
Drug sale	0.005	0.116	0.180
Drug possession	0.000	0.929	0.610

Notes: This table reports the results of the statistical tests using Stevenson (2018)'data, including time fixed effects as controls. Specifically, the considered controls are year, month, and day of the week fixed effects. *Sharp Test* stands for our novel semi-parametric test developed in this paper, while *FLL-Nonp* and *FLL-Semip* represent the nonparametric and semi-parametric tests of FLL (three knots B-spline), respectively.

tions to salvage the model. In this section, we consider partial exclusion and monotonicity to weaken the assumptions. Consider a general form of the potential outcome model:

$$Y = \tilde{Y}_1 D + \tilde{Y}_0(1-D), \quad \tilde{Y}_d = \sum_{z\in\mathcal{Z}} Y_d(z)1\{Z=z\}, \quad D = \sum_{z\in\mathcal{Z}} D_z 1\{Z=z\}.$$

Let us decompose Z into two components $Z \equiv (Z_I, Z_c)$ where Z_I is the judge fixed effect, and Z_c is a vector of judge observable characteristics, such as experience, race, and political party, among others.

Assumption 5.1 ***(Partial Exclusion Restriction)*** *For $d \in \{0,1\}$, $Y_d(z) = Y_d(z_c)$ for all $z \in \mathcal{Z}$.*

This partial exclusion assumption relaxes Assumption 2.2. It allows the potential outcomes to depend on judges' observable characteristics but not their fixed effects or identity. For instance, when the treatment of interest is incarceration, judges could assign and differ in other punishments like probation, fines, or sentence length. These other punishments could directly affect potential outcomes, making Assumption 2.2 unlikely. For example, minority judges may be less lenient in their sentence length than their majority counterparts (Johnson, 2014). Beyond the decision to incarcerate, different sentence lengths may have divergent effects on, for example, later labor market outcomes. If the sentence length is not observed or controlled, we would expect the potential outcome to depend on whether a judge is a minority judge through this channel. The partial exclusion assumption says whether or how the judge assigns other types of punishment only depends on the judge's observable (Z_c), but not the judge's identity (Z_I). In other words, a defendant will end up with the same pair of potential outcomes $(Y_1(Z_c), Y_0(Z_c))$ as long as he or she is assigned to judges with the same observed Z_c. Finally, when the only instrument variable we observe in the data is the identity of the judge Z_I, then the partial exclusion assumption is equivalent to the original exclusion Assumption 2.2. The next assumption relaxes the monotonicity condition.

Assumption 5.2 (Partial Monotonicity) *For any (z_I, z_c) and $(z'_I, z_c) \in \mathcal{Z} \times \mathcal{Z}$ either $D(z_I, z_c) \geq D(z'_I, z_c)$ for all defendants or $D(z_I, z_c) \leq D(z'_I, z_c)$ for all defendants.*

Mogstad, Torgovitsky, and Walters (2019) introduced partial monotonicity assumption, which significantly weakens Assumption 2.3 since it does not require comparing the level of leniency across judges with different observable characteristics. For instance, let $Z_c = (Z_c^R, Z_c^P)$ be composed of the following binary variables: Z_c^R equal to 1 if the judge is black or Hispanic and 0 if not, while Z_c^P is 1 if the judge is from the Republican party and 0 if from the Democratic party. Imposing Assumption 2.3 means it is not possible to have a black democrat judge be more lenient than a white republican judge for some defendants while being less lenient for other defendants, i.e. these two judges may have different cut-off points, but they rank all the defendants by the same order. Mathmatically, we can not have both $\mathbb{P}(1 = D(z_I, 1, 0) > D(z'_I, 0, 1) = 0) > 0$ and $\mathbb{P}(1 = D(z'_I, 0, 1) > D(z_I, 1, 0) = 0) > 0$. However, there is a large empirical evidence of heterogeneity in the ranking of judges' leniency across different types of offence or defendants (see Abrams, Bertrand, and Mullainathan, 2012; Stevenson, 2018). This is, however, compatible with the partial monotonicity. Its main advantage is it no longer requires a uniform ranking of defendants across different judges. Judges' rankings are allowed to vary with their characteristics Z_c.

Assumption 5.3 (Partial Single Threshold-Crossing) *Type $Z = (Z_I, Z_c)$'s judge treatment assignment mechanism is governed by the following threshold crossing model $D = 1\{\nu(Z_I, Z_c) \geq U_{Z_c}\}$ for a measurable function ν, where the distribution of U_{z_c} is absolutely continuous for all $z_c \in \mathcal{Z}_c$.*

Under Assumptions 2.1 and 5.3, we can rewrite the partial STC without loss of generality as follows:

$$D(z_I, z_c) = 1\left\{F_{U_{z_c}}(\nu(z_I, z_c)) \geq F_{U_{z_c}}(U_{z_c})\right\} \equiv 1\left\{P(z_I, z_c) \geq V_{z_c}\right\},$$

where $F_{U_{z_c}}(\cdot)$ is the distribution function of U_{z_c}, $P(z_I, z_c)$ is identified from the observed (D, Z) by $P(z_I, z_c) \equiv \mathbb{P}(D = 1|Z_I = z_I, Z_c = z_c)$. Note by construction, V_{z_c} follows $Uniform[0, 1]$ distribution because the distribution of U_{z_c} is absolute continuous; also, V_{z_c} is independent with (Z_I, Z_C).

The key difference between the STC and the Partial STC is even though V_{z_c} follows $Uniform[0, 1]$ distribution, each defendant does not have a single V. Instead, he or she faces a collection of $\{V_{z_c}, z_c \in \mathcal{Z}_c\}$. This unobserved latent variable is now different for judges with distinct observable characteristics. The partial STC has a natural interpretation as an extension of the Roy model (Canay, Mogstad, and Mountjoy, 2020). We can interpret $P(z_I, z_c)$ as the perceived gain of incarcerating a defendant by a type $z = (z_I, z_c)$ judge, and V_{z_c} as the expected cost (but unobserved to the econometrician) of an incarcerating a defendant. The particularity of the partial STC is the expected cost can vary across judges with distinct observable characteristics z_c but fixed within judges with the same z_c. In standard monotonicity assumption, the cost V would be the same regardless of the characteristics (z_I, z_c).

The partial STC has an alternative interpretation, allowing decision-makers (judges) to differ in their preferences and skills. Chan, Gentzkow, and Yu (2022) argue the standard STC (Assumption 2.4) implies the data must be consistent with all judges having the same signal U (the same skill). However, under the partial STC, judges with distinct Z_c are allowed to differ in their signal U_{z_c} and thus, in their skill. In consequence, Assumption 5.3 significantly enhances the pattern of heterogeneity between judges in the judge leniency design.

Let us consider a situation where eight judges decide whether to incarcerate a given defendant to further elucidate the richer heterogeneity enabled by the partial monotonicity (or equivalently, the partial STC) assumption. We consider the two observable characteristics of the judges introduced earlier, $Z_c \equiv (Z_c^R, Z_c^P) \in \{0, 1\} \times \{0, 1\}$. These two binary observable characteristics lead to four judges' types. The eight judges are evenly allocated across these four types.

The left rectangle of Figure 6 shows the benefit and the expected cost of incarcerating

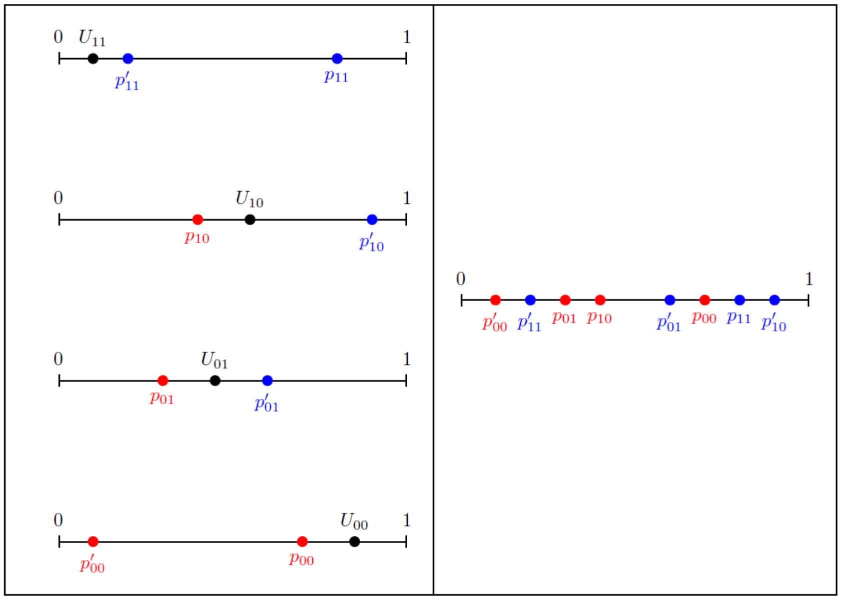


Figure 6: Monotonicity in Judge IV and Conditional Judge IV designs

the defendant on a separate unit segment for each judge. For example, $p_{11} and \mathrm{p}'_{11}$ are the benefit of the two black democratic judges with type $Z_C = (1, 1)$ to incarcerate the defendant. The right rectangle of Figure 6 plots the benefit numbers of all eight judges on the same unit segment. Similarly, U_{11} represents the expected cost of incarcerating the defendant by a black democratic judge: they share the same expected cost or skills. A judge incarcerates the defendant when the corresponding benefit is higher than the expected cost of incarceration. In Figure 6, the judges who incarcerate the defendant are blue-coloured, while those of judges who release the defendant are red-coloured.

The behaviour of the eight judges does not violate Assumption 5.2 or Assumption 5.3. However, the standard monotonicity Assumption 2.3 is clearly violated (right rectangle of Figure 6). Indeed, the judge with benefit score p'_{11} incarcerates the defendant (blue-coloured), whereas judges with higher benefit scores p_{01}, p_{10}, or p_{00} do not incarcerate the defendant (red-coloured). Note that Assumption 2.3 would not be violated for this group of judges only under one of these two conditions: (i) all four V_{z_c} are greater than the maximum of the eight propensities or smaller than the minimum of all eight properties. In other words, when all judges make the same decision regarding this defendant, or (ii) judges who do not incarcerate the defendant must have lower benefit scores than judges who incarcerate the defendant. Moreover, one of these two conditions must hold for all defendants when we impose Assumption 2.3 (or Assumption 2.4).

However, Assumption 5.2 (or Assumption 5.3) does not require such a binding restriction. In particular, under the partial monotonicity assumption, defendants are allowed to be defiers across judges with distinct observed characteristics Z_C. For instance, in Figure 6 and using propensity scores to identify judges, the defendant is a $p'_{11} - p_{01}$ defier, a $p'_{11} - p_{10}$ defier, a $p'_{11} - p_{00}$ defier, and a $p'_{01} - p_{00}$ defier.

Assumptions 2.1, 5.1 and 5.3 are weaker than Assumptions 2.1 to 2.3. We show that under these weaker conditions, it is still possible to identify meaningful treatment effect parameters.

Theorem 3 (Identification of causal effects under Partial exclusion and monotonicity) *If Assumptions 2.1, 5.1 and 5.3 hold, then:*

(i) (Identification of the LATE). Let $\mathcal{P}_{z_c}$ be the support of $P(Z_I, Z_c)$ conditioning on $Z_c = z_c$. Then for any pair $(p_{z_c}, p'_{z_c}) \in \mathcal{P}_{z_c} \times \mathcal{P}_{z_c}$ such that $p_{z_c} < p'_{z_c}$ we have the following identification results:

$$\frac{\mathbb{E}[g(Y)|P = p'_{z_c}, Z_c = z_c] - \mathbb{E}[g(Y)|P = p_{z_c}, Z_c = z_c]}{p'_{z_c} - p_{z_c}} = \mathbb{E}[g(Y_1(z_c)) - g(Y_0(z_c))|1\{p_{z_c} < V_{z_c} \leq p'_{z_c}\}].$$

(ii) (Identification of the MTE). For any $p_{z_c} \in \mathcal{P}_{Z_c}$ such that $\mathbb{E}[g(Y)|P = \cdot, Z_c = z_c]$ is continuously differentiable in the neighborhood of p_{z_c}, then,

$$\frac{\partial \mathbb{E}[g(Y)|P = t, Z_c = z_c]}{\partial t}\Big|_{t=p_{z_c}} = \mathbb{E}[g(Y_1(z_c)) - g(Y_0(z_c))|V_{z_c} = p_{z_c}].$$

(iii) (Testable restrictions). For any fixed $z_c \in \mathcal{Z}_c$, $\mathbb{P}(y < Y \leq y', D = 1|P = p, Z_c = z_c)$ and $-\mathbb{P}(y < Y \leq y', D = 0|P = p, Z_c = z_c)$ are non-decreasing in p for all $p \in \mathcal{P}_{Z_c}$ and any $y, y' \in \mathcal{Y}$.

The proof of Theorem 3 is similar to Theorem 1 after conditioning on $Z_c = z_c$ and therefore omitted. The identification results stated in Theorem 3 (i)-(ii) demonstrate whenever there are two judges with distinct Z_I but share the same observed characteristics $Z_c = z_c$, the conditional Wald estimand identifies the LATE provided the propensity scores for these two judges are different. Moreover, when the distribution of $Z_I|Z_c = z_c$

allows one to take the derivative of $\mathbb{E}[g(Y)|P = \cdot, Z_c = z_c]$, the conditional LIV estimand identifies the MTE. This identification result is a local version of the standard LATE and MTE identification.

Theorem 3 (iii) presents the testable implications of the weaker monotonicity and exclusion assumptions. The testable implications in Theorem 3 (iii) are weaker than those in Theorem 1 (i). To see this, let us consider the same example of the eight judges discussed above, where the outcome of interest is recidivism ($Y \in \{0,1\}$). We consider the same two observable characteristics of the judges, $Z_c \equiv (Z_c^R, Z_c^P) \in \{0,1\} \times \{0,1\}$. Let $\theta^d(p) = \mathbb{P}(Y = 0, D = d|P = p)$ for $d \in \{0,1\}$. In this simple case, the sharp testable implications under the standard judge leniency design, i.e. Assumptions 2.1 to 2.3 are:

$$\theta^1(p'_{00}) \leq \theta^1(p'_{11}) \leq \theta^1(p_{01}) \leq \theta^1(p_{10}) \leq \theta^1(p'_{01}) \leq \theta^1(p_{00}) \leq \theta^1(p_{11}) \leq \theta^1(p'_{10})$$
$$\theta^0(p'_{00}) \geq \theta^0(p'_{11}) \geq \theta^0(p_{01}) \geq \theta^0(p_{10}) \geq \theta^0(p'_{01}) \geq \theta^0(p_{00}) \geq \theta^0(p_{11}) \geq \theta^0(p'_{10}),$$

which is a total of fourteen inequalities. However, when invoking our weaker set of assumptions, we have only eight inequalities which characterize the sharp testable implications:

$$\theta^1(p'_{11}) \leq \theta^1(p_{11}), \quad \theta^1(p_{10}) \leq \theta^1(p'_{10}), \quad \theta^1(p_{01}) \leq \theta^1(p'_{01}), \quad \theta^1(p'_{00}) \leq \theta^1(p_{00})$$
$$\theta^0(p'_{11}) \geq \theta^0(p_{11}), \quad \theta^0(p_{10}) \geq \theta^0(p'_{10}), \quad \theta^0(p_{01}) \geq \theta^0(p'_{01}), \quad \theta^0(p'_{00}) \geq \theta^0(p_{00}).$$

The comparison of the testable implications in Theorems 1 and 3 confirms the judge leniency design is more stringent than the conditional judge leniency design. Hence, whenever the standard judge leniency design is rejected, the researcher may rely on its relaxed versions as long as the testable implications derived in Theorem 3 are satisfied.

6 Conclusion

In this paper, we derive the sharp testable implications for identifying assumptions for the judge's leniency design in a general framework where the instruments can be either discrete or continuous and propose a consistent test for the implications. Our simulation study and empirical results identify how essential it is to consider sharp implications for the best use of information in the data. While we focus on the primary application of testing the validity of judge leniency design, our method can readily be applied to a broad range of applications.

APPENDIX

A Constructing $\hat{\nu}_d^b(\ell)$

In this appendix, we show how to construct the bootstrap estimates $\hat{\nu}_d^b(\ell)$. For bootstrap iteration b, let $\{W_1^b, W_2^b, \cdots, W_n^b\}$ be a sequence of i.i.d. random variables with zero mean and unit variance. For instance, we can choose standard normal. Let $\hat{\theta}^b$ be the MLE based on the b-th bootstrapped sample:

$$\begin{aligned}
\hat{\theta}^b &= \underset{\theta \in \Theta}{\operatorname{argmax}} \frac{1}{n} \sum_{i=1}^{n} W_i^b \log f(Y_i, D_i; Z_i\theta) \\
&\equiv \underset{\theta \in \Theta}{\operatorname{argmax}} \frac{1}{n} \sum_{i=1}^{n} W_i^b \left\{ D_i \log P(Z_i, \theta) + (1 - D_i) \log(1 - P(Z_i, \theta)) \right\}. \quad \text{(A.1)}
\end{aligned}$$

We define the weighted bootstrapped estimators for $m_1(y, r_y, p, r_p, \theta_0)$, $m_0(y, r_y, p, r_p, \theta_0)$ and $w(p, r_p, \theta_0)$ be

$$\begin{aligned}
\widehat{m}_1^b(y, r_y, p, r_p, \hat{\theta}^b) &= \frac{1}{n} \sum_{i=1}^{n} W_i^b \cdot m_{1i}(y, r_y, p, r_p, \hat{\theta}^b) \Big/ \frac{1}{n} \sum_{i=1}^{n} W_i^b, \\
\widehat{m}_0^b(y, r_y, p, r_p, \hat{\theta}^b) &= \frac{1}{n} \sum_{i=1}^{n} W_i^b \cdot m_{0i}(y, r_y, p, r_p, \hat{\theta}^b) \Big/ \frac{1}{n} \sum_{i=1}^{n} W_i^b, \\
\widehat{w}^b(p, r_p, \hat{\theta}^b) &= \frac{1}{n} \sum_{i=1}^{n} W_i^b \cdot w_i(p, r_p, \hat{\theta}^b) \Big/ \frac{1}{n} \sum_{i=1}^{n} W_i^b,
\end{aligned}$$

Finally, for a given $\ell = (y, r_y, p_1, p_2, r_p)'$, we can construct $\hat{\nu}_d^b(\ell)$ for the b-th bootstrap iteration

$$\hat{\nu}_1^b(\ell) = \hat{m}_1^b(y, r_y, p_2, r_p, \hat{\theta}) \cdot \hat{w}^b(p_1, r_p, \hat{\theta}^b) - \hat{m}_1^b(y, r_y, p_1, r_p, \hat{\theta}^b) \cdot \hat{w}^b(p_2, r_p, \hat{\theta}^b), \quad \text{(A.2)}$$

$$\hat{\nu}_0^b(\ell) = \hat{m}_0^b(y, r_y, p_2, r_p, \hat{\theta}^b) \cdot \hat{w}^b(p_1, r_p, \hat{\theta}^b) - \hat{m}_0^b(y, r_y, p_1, r_p, \hat{\theta}^b) \cdot \hat{w}^b(p_2, r_p, \hat{\theta}^b). \quad \text{(A.3)}$$

References

Abrams, D. S., M. Bertrand, and S. Mullainathan (2012): "Do judges vary in their treatment of race?," *The Journal of Legal Studies*, 41(2), 347–383.

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

Andrews, D. W. K., and X. Shi (2013): "Inference Based on Conditional Moment Inequalities," *Econometrica*, 81(2), 609–666.

Bhuller, M., G. B. Dahl, K. V. Løken, and M. Mogstad (2018): "Incarceration spillovers in criminal and family networks," Discussion paper, National Bureau of Economic Research.

Brinch, C. N., M. Mogstad, and M. Wiswall (2017): "Beyond LATE with a Discrete Instrument," *Journal of Political Economy*, 125(4), 985–1039.

Canay, I. A., M. Mogstad, and J. Mountjoy (2020): "On the use of outcome tests for detecting bias in decision making," Discussion paper, National Bureau of Economic Research.

Carneiro, P., J. J. Heckman, and E. Vytlacil (2010): "Evaluating marginal policy changes and the average effect of treatment for individuals at the margin," *Econometrica*, 78(1), 377–394.

Carneiro, P., J. J. Heckman, and E. J. Vytlacil (2011): "Estimating marginal returns to education," *American Economic Review*, 101(6), 2754–2781.

Carneiro, P., and S. S. Lee (2009): "Estimating distributions of potential outcomes using local instrumental variables with an application to changes in college enrollment and wage inequality," *Journal of Econometrics*, 149(2), 191–208.

Carr, T., and T. Kitagawa (2021): "Testing instrument validity with covariates," *arXiv preprint arXiv:2112.08092*.

Chan, D. C., M. Gentzkow, and C. Yu (2022): "Selection with variation in diagnostic skill: Evidence from radiologists," *The Quarterly Journal of Economics*, 137(2), 729–783.

Cunningham, S. (2021): *Causal inference: The mixtape.* Yale university press.

Di Tella, R., and E. Schargrodsky (2013): "Criminal recidivism after prison and electronic monitoring," *Journal of Political Economy*, 121(1), 28–73.

Dobbie, W., P. Goldsmith-Pinkham, and C. S. Yang (2017): "Consumer bankruptcy and financial health," *Review of Economics and Statistics*, 99(5), 853–869.

Dobbie, W., H. Grönqvist, S. Niknami, M. Palme, and M. Priks (2018): "The intergenerational effects of parental incarceration," Discussion paper, National Bureau of Economic Research.

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

Farre-Mensa, J., D. Hegde, and A. Ljungqvist (2020): "What is a patent worth? Evidence from the US patent "lottery"," *The Journal of Finance*, 75(2), 639–682.

Frandsen, B., L. Lefgren, and E. Leslie (2023): "Judging Judge Fixed Effects," *American Economic Review*, 113(1), 253–77.

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

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

Holm, S. (1979): "A simple sequentially rejective multiple test procedure," *Scandinavian journal of statistics*, pp. 65–70.

Hsu, Y.-C. (2017): "Consistent tests for conditional treatment effects," *The econometrics journal*, 20(1), 1–22.

Hsu, Y.-C., and R. P. Lieli (2021): “Inference for ROC curves based on estimated predictive indices,” *arXiv preprint arXiv:2112.01772*.

Hsu, Y.-C., C.-A. Liu, and X. Shi (2019): “Testing generalized regression monotonicity,” *Econometric Theory*, 35(6), 1146–1200.

Huber, M., and G. Mellace (2015): “Testing instrument validity for LATE identification based on inequality moment constraints,” *Review of Economics and Statistics*, 97(2), 398–411.

Johnson, B. D. (2014): “Judges on trial: A reexamination of judicial race and gender effects across modes of conviction,” *Criminal Justice Policy Review*, 25(2), 159–184.

Kédagni, D., L. Li, and I. Mourifié (2020): “Discordant relaxations of misspecified models,” *arXiv preprint arXiv:2012.11679*.

Kitagawa, T. (2015): “A Test for Instrument Validity,” *Econometrica*, 83(5), 2043–2063.

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

Kowalski, A. E. (2016): “Doing more when you’re running late: Applying marginal treatment effect methods to examine treatment effect heterogeneity in experiments,” Discussion paper, National Bureau of Economic Research.

Mogstad, M., A. Torgovitsky, and C. Walters (2019): “Identification of causal effects with multiple instruments: Problems and some solutions,” *NBER Working Paper*, (w25691).

Mourifié, I., and Y. Wan (2017): “Testing Local Average Treatment Effect Assumptions,” *The Review of Economics and Statistics*, 99(2), 305–313.

Mueller-Smith, M. (2015): “The criminal and labor market impacts of incarceration,” *Unpublished Working Paper*, 18.

Norris, S., M. Pecenco, and J. Weaver (2021): “The effects of parental and sibling incarceration: Evidence from ohio,” *American Economic Review*, 111(9), 2926–2963.

Stevenson, M. T. (2018): "Distortion of justice: How the inability to pay bail affects case outcomes," *The Journal of Law, Economics, and Organization*, 34(4), 511–542.

Sun, Z. (2023): "Instrument validity for heterogeneous causal effects," *Journal of Econometrics*, 237(2), 105523.

Vytlacil, E. (2002): "Independence, Monotonicity, and Latent Index Models: An Equivalence Result," *Econometrica*, 70(1), 331–341.