A Sharp Test for the Judge Leniency Design*

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

August 4, 2025

Abstract

We propose sharp testable implications and tests to jointly assess the random assignment, exclusion, and monotonicity assumptions in judge leniency designs. Our procedures accommodate various data scenarios in which the number of defendants handled by a judge may be either small or large, and allow for discrete or continuous instrumental variables. When the validity of the design is rejected, a variant of the marginal treatment effect can be identified under weaker assumptions. We apply our test to the Philadelphia court data studied by Stevenson (2018) and demonstrate that it outperforms non-sharp joint 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 SSHRC Insight Grant #43520190500. Wan thanks the SSHRC Insight Grant #43520240428. 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, such as pretrial detentions and incarcerations, on subsequent outcomes, including 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. This means that if a defendant was treated (detained) by one judge, she would always be treated (detained) by a less lenient judge. Trial decisions (treatment) are often multidimensional, including incarceration, fines, community service, sentence length, and others (Johnson, 2014). These decisions impact future outcomes. Because different judges may have varying attitudes on these decisions, the exclusion restriction can be violated if some of the decisions are un-

[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.

observed 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 offenses. 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. These observations align with Mogstad, Torgovitsky, and Walters (2019), who demonstrate that, in general, monotonicity effectively requires homogeneous choice behavior for economic agents when there are multiple instruments. Therefore, offering a statistical test to evaluate the validity of the judge leniency design becomes a highly relevant empirical question.

In this paper, we characterize the *sharp testable implications* of the judge leniency design as a set of inequality restrictions on the distribution of the observed data. Our result is novel and contributes to the testable implications derived in the seminal work of Heckman and Vytlacil (2005) in two ways. First, our implications belong to a tractable subset of the constraints of Heckman and Vytlacil (2005) and are easier to implement in practice. Second, we establish the sharpness of our testable implication, that is, they 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.

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) examines 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. Our approach adds to the existing body of knowledge by introducing a test that does not depend on prior information. 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) proposes 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 significant 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 numerical examples and empirical studies reported in Sections 2 and 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 Li, Kédagni, and Mourifié, 2024). 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 literature.

We construct valid and consistent semi-nonparametric and semiparametric tests based on these tractable testable implications. Our asymptotic tests support a diverse range of data structures. For example, we can apply our tests to the empirical context in which each judge handles a large number of defendants, and the number of judges can be either large or small (as in our empirical application). As we will further elaborate in Section 3, our asymptotic tests are also applicable when the number of defendants for each judge is small, as long as the data regime permits a root-n estimation of the propensity score.

We also provide an easy-to-compute finite sample test for cases involving a small number of judges and a small number of defendants per judge. To the best of our knowledge, ours and the finite sample test of Frandsen, Lefgren, and Leslie (2023) are the only finite sample specification tests in the judge leniency design literature. Like Frandsen, Lefgren, and Leslie (2023), our finite sample test also focuses on binary outcomes. Unlike Frandsen, Lefgren, and Leslie (2023)'s test, which ensures the finite sample validity by computing a "least favorable p-value" via a high-dimensional nonlinear optimization routine, we use Bonferroni correction. The computation for our test is very light and requires little more than simulating Bernoulli random variables. Therefore, it serves as a useful addition to the existing finite sample 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 Treatment Effect (LATE) framework, i.e., Kitagawa (2015), Huber and Mellace (2015), and Mourifié and Wan (2017). However, it is worth noting that 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 considering 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.

While our test is primarily motivated by testing judge leniency designs, it can also be applied to assess 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. Finally, motivated by Mogstad, Torgovitsky,

and Walters (2019), we propose to relax monotonicity and exclusion assumptions to partial monotonicity and partial exclusion, respectively, when our test rejects the null hypothesis.

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, the unit of observation is defendants. Hence, $D = 1$ indicates that a defendant is incarcerated. Let $Z \in \mathcal{Z} \subseteq \mathbb{R}^{d_z}$ be the type of the judge assigned to the defendant. $Y_d(z) \in \mathcal{Y} \subseteq \mathbb{R}$ denotes the potential outcome of interest (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, continuous, discrete, or a combination of both. For example, if there is a group of judges $\mathcal{J}$, and if 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 that judges are usually randomly assigned to different cases, making the random assignment assumption likely to hold in practice. 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 that any defendants treated (incarcerated) by a more lenient judge be also treated if assigned to a less lenient one. Heckman and Vytlacil (2005) refers to the monotonicity assumption as a uniformity condition since it restricts that the treatment on all the defendants must vary in a uniform direction when externally assigned to another 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:

$$\begin{cases} Y = Y_1 D + Y_0(1-D), \\ D = 1\{P(Z) \geq V\}. \end{cases} \tag{2.1}$$

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, we will discuss 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 Appendix B.1. 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 subclass 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.[3]

The second part of Theorem 1 is new, and it shows that 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 Li, Kédagni, and Mourifié (2024), 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 may 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.

[3]We note that use the the half-interval class $g(Y) = 1\{Y \leq y\}, y \in \mathcal{Y}$ will result in loss of power. To see this, suppose the support is finite, that is, $\mathcal{Y} = \{y_1, y_2, \cdots, y_K\}$, then it is without loss of information to consider the class of singletons: $g(Y) = 1\{Y = y_k\}$, $k = 1, 2, \cdots, K$. However, if one considers $g(Y) = 1\{Y \leq y_k\}$, then it is possible that both $\mathbb{P}(Y \leq y_1, D = 1|P = p)$ and $\mathbb{P}(Y \leq y_2, D = 1|P = p)$ are non-decreasing function, but $\mathbb{P}(Y = y_2, D = 1|P = p)$ is not.

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 LIV 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 an 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, suppose $\mathcal{Z} = \{z_1, ..., z_K\}$ 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. Given this imposed ordering, Assumptions 2.1

to 2.3 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 that 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 that 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 non-negligible 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:

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

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:

$$L_g - U_g \leq W(g(Y), p, p') \leq U_g - L_g, \tag{2.2}$$

where the inequality in (2.2) 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:

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

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

$$L_g \leq W(g(YD), p, p') \leq U_g, \tag{2.3}$$
$$-U_g \leq W(g(Y(1-D)), p, p') \leq -L_g, \tag{2.4}$$

One can easily observe that the testable restrictions in (2.3) and (2.4) could be violated, whereas the restriction used by FLL, i.e. inequality (2.2) still holds. Hence, implementing FLL's statistical testing procedure based on inequalities (2.3) or (2.4) could provide a different result compared to their test based on inequality (2.2) alone. These discordant implications confirm the concern about developing a statistical test based on non-sharp restrictions. Example 2.1 provides a concrete numerical example.

Example 2.1 *Consider the potential outcome model:*

$$\begin{cases} Y = Y_1 D + Y_0(1-D), \\ D = 1\{P \geq V\}. \end{cases}$$

Suppose V is independent of (Y_1, Y_0, P). However, (Y_1, Y_0) and P are dependent:

$$\begin{cases} Y_1|P = \tilde{p} \sim \text{ degenerate at 1}, & \text{if } \tilde{p} < \frac{1}{2} \\ Y_1|P = \tilde{p} \sim Bernoulli(\tilde{p}), & \text{if } \tilde{p} \geq \frac{1}{2} \end{cases}$$

$$\begin{cases} Y_0|P=\tilde{p} \sim \ \textit{degenerate at 0}\ , & \textit{if } \tilde{p} < \frac{1}{2} \\ Y_0|P=\tilde{p} \sim \textit{Bernoulli}(\tilde{p}), & \textit{if } \tilde{p} \geq \frac{1}{2} \end{cases}$$

Therefore, the randomization assumption is violated, but the monotonicity and exclusion conditions are met. In this case, Y_d is binary and $U_g = 1$ and $L_g = 0$, so we also take $g(\cdot)$ to be the identity function without loss of generality.

For this DGP, we can show that for any $p' \in (0,1)$ and $p \in (0,1)$, $W(g(Y), p, p') = 1$. Hence, FLL's testable implication (inequality 2.2 above) always holds and has no power to detect the violation. There is missing information. For example, when $p' > p > \frac{1}{2}$, we can verify that $W(g(YD), p, p') = p' + p > 1 \equiv U_g$. Therefore, condition (2.3) is violated. Please see derivation details in Appendix B.4.

On the other hand, our testable implication can capture such a violation. To see this, note

$$\mathbb{E}[YD|P=p] = \mathbb{E}[Y_1|P=p]p = \begin{cases} p & \textit{if } p < \frac{1}{2}, \\ p^2 & \textit{if } p \geq \frac{1}{2}. \end{cases} .$$

It is apparent that $\mathbb{E}[YD|P=p]$ is not a monotone function of p, and therefore violates our testable implication. □

The intuition behind Example 2.1 is not pathological and is reflected in the derivation in Appendix B.4. Because $\mathbb{E}[Y|P=p] = \mathbb{E}[Y_1 D|P=p] + \mathbb{E}[Y_0(1-D)|P=p]$, it is possible the violations on the $\mathbb{E}[Y_1 D|P=p]$ and $\mathbb{E}[Y_0(1-D)|P=p]$ "cancel" out. As a consequence, the quantity $\mathbb{E}[Y|P=p]$ provides no power to detect violations in these cases.

Another evident reason why FLL's implications cannot exhaust all violations of the judge leniency design is that they only focus on $g(Y) = Y$, whereas the inequality in (2.2) 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 (2.2), 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 (2.2) holds trivially and FLL's test 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 impli-

cations 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

In this section, we construct tests based on Theorem 1. For the defendant $i \in \{1, 2, \cdots, n\}$, researchers observed a vector (Y_i, D_i, Z_i, X_i), where Y_i, D_i, Z_i, and X_i represent his/her observed outcome, observed treatment status, the vector of characteristics of the judge that i was assigned to, and the vector of additional control variables, respectively.

We first present our baseline semi-nonparametric test in Section 3.1 without the presence of control variables X_i. For this test, we make no functional form or distributional assumptions about potential outcomes. We do need to estimate the propensity score $P(z) \equiv \mathbb{P}(D_i = 1|Z_i = z)$ first, for which our procedure can accommodate different data scenarios. If Z_i contains continuous variables, 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).[4] When Z_i only contains discrete variables, such as judge's gender, we can estimate $\mathbb{P}(D_i = 1|Z_i = z)$ by the sample averages of D conditioning on each possible value of $z \in \mathcal{Z}$. In this case, our test is indeed nonparametric.

We should emphasize that, in both cases above, we do not require any knowledge of the identity of judges, nor do we need the number of defendants handled by each judge to diverge to infinity. For example, when Z is gender, we only need the number of defendants for each judge's gender to go to infinity. This can happen when the number of judges is large, but each judge handles a finite number (or even one) of defendants. Therefore, our test can also be applied to other empirical contexts than the judge leniency design. There is another scenario in which Z_i is the judge's identity. Suppose the number of defendants handled by each judge is large, as in our empirical application. In this case, we can also

[4] 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.

consistently estimate judge j's propensity score $\mathbb{P}(D=1|Z_i=j)$ by the sample frequency estimator for each judge j, regardless of whether the number of judges is small or large.

In practice, the number of defendants that a judge handles can be small. In this case, one can not estimate $\mathbb{P}(D=1|Z_i=j)$ consistently without additional assumptions and conventional inference methods can be invalid. This phenomenon has received attention from the literature, see discussions in Jochmans (2023), Ren (2024), Sithole (2024), and Yap (2024). These papers, however, focus on inference on the parameters instead of testing model specification. To account for data scenarios with small numbers of defendants per judge, we design a test that does not require a consistent estimator for the propensity score and controls the size at any sample sizes for the case of a binary outcome. To the best of our knowledge, this test and FLL's finite sample test are the only ones for testing judge leniency design specification with finite samples, and both focus on binary outcome variables. Our test uses upper bounds of the null distribution to calculate the critical values, and hence is very easy to implement. It only requires simulating Bernoulli random variables, and no nonlinear optimization is involved. For the purpose of exposition, we collect the finite sample test in Appendix C and focus on the cases in which the propensity score can be consistently estimated in this section.

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.

3.1 A Semi-nonparametric test

For the convenience of the exposition, we restate 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 $[0,1]$.[5] 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 that the conditioning variable P is not directly observed from the data. We derive the new influence functions and show that 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:

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

and

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

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

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

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:

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

[5] 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 re-centering: $\tilde{Y} = \Phi\left(\frac{Y - \bar{Y}}{\hat{std}(Y)}\right)$, where $\bar{Y}$ is the sample average and $\hat{std}(Y)$ is the sample standard deviation.

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 $P(Z_i)$ is unknown, but can be replaced by its root-n consistent estimate $\hat{P}_i$. When we estimate the propensity score by a parametric model, we denote it as $\hat{P}_i \equiv P(Z_i, \hat{\theta})$, where $\hat{\theta}$ is the MLE. When Z_i is the judge's identity and the number of defendants for each judge is large, we simply use the frequency estimator $\hat{P}_i = \frac{\sum_{j=1}^n D_j 1\{Z_j = Z_i\}}{\sum_{j=1}^n 1\{Z_j = Z_i\}}$. 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. *Compute the estimator for the propensity score $\hat{P}_i$, as detailed in Appendix A.*

3. *For each $\ell \in \mathcal{L}_Q$, construct estimates $\hat{\nu}_1(\ell)$ and $\hat{\nu}_0(\ell)$ as sample analogs of Equations (3.3) and (3.4), as detailed in Equations (A.5) and (A.6).*

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

   (a) *Draw $W_1^b, W_2^b, \cdots, W_n^b$ 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 $\hat{P}_i^b$, defined in Equation (A.9).*

   (c) *Obtain $\hat{\nu}_d^b(\ell)$, $d = 0, 1$, for each bootstrap iteration using Equations (A.10) and (A.11).*

5. *Compute the normalization factor, denoted by* $\hat{\sigma}_d(\ell)$*, as:*

$$\hat{\sigma}_d^2(\ell) = \frac{n}{B}\sum_{b=1}^{B}\left(\hat{\nu}_d^b(\ell) - \overline{\hat{\nu}}_d^b(\ell)\right)^2, \text{ where } \overline{\hat{\nu}}_d^b(\ell) = \frac{1}{B}\sum_{b=1}^{B}\hat{\nu}_d^b(\ell). \tag{3.6}$$

Choose a constant $\epsilon > 0$*, and let* $\hat{\sigma}_{d,\epsilon}^2(\ell) = \max\{\hat{\sigma}_d^2(\ell), \epsilon\}$*.*

6. *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*

$$\widehat{T}_n = \sum_{d=0,1}\sum_{\ell\in\mathcal{L}_Q}\max\left\{\sqrt{n}\frac{\hat{\nu}_d(\ell)}{\hat{\sigma}_{d,\epsilon}(\ell)}, 0\right\}^2\Omega(\ell).[6] \tag{3.7}$$

7. *Let* a_n *and* B_n *be positive deterministic sequences.*[7] *Calculate the generalized moment selection (GMS) terms as*

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

8. *For* $b = 1, 2, \cdots, B$*, calculate the quantity*

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

where

$$\Phi_d^b(\ell) = \sqrt{n}\left(\hat{\nu}_d^b(\ell) - \hat{\nu}_d(\ell)\right). \tag{3.8}$$

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.* $\eta = 10^{-6}$*.*[8]

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

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

[7] 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] η is the 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}.

Theorem 2 shows that the test ϕ_n has its size controlled asymptotically and is consistent. The proof for Theorem 2 is collected in Appendix B.2 of the online supplementary material. We also list all the technical assumptions, such as conditions that ensure the first-stage estimator converges at a sufficiently fast rate, in that section for the sake of exposition.

Theorem 2 *Suppose Assumptions B.1 to B.3 and B.5 in Appendix B.2 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,

$$\limsup_{n\to\infty} \mathbb{P}(\phi_n = 1|H_1) = 1. \tag{3.10}$$

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 $D = 1\{\nu(Z, X) \geq U\}$ 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 $p_1 \geq p_2$, and all $y, y' \in \mathcal{Y}$

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

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

Remark 3.1 *If X is discrete and $\mathcal{X}$ only contains a relatively small number of values, or X contains a small number of continuous variables, we can also follow the same procedure as in Section 3.1 but add cubes for X. Under the null hypothesis, we should expect $T \equiv \sum_{d=0,1} \sum_{x \in \mathcal{X}} \sum_{\ell \in \mathcal{L}} \max\{\nu_d(\ell, x), 0\}^2 \Omega(\ell, x) = 0$, where $\Omega(\ell, x)$ is a positive weighting function chosen by researchers, and*

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

and

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

and r_x is similarly defined as r_p and r_y. The implementation follows analogously 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

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

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

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 that 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 $\tilde{Y}$. Assumption 3.3 is a common assumption made in the literature for estimating the MTE, see for instance Carneiro and Lee (2009); Carneiro, Heckman, and Vytlacil (2010); Kowalski (2016). Nevertheless, we do acknowledge it is subject to the potential risk of model mis-specification. 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 $D = 0$:

$$\mathbb{E}[Y|X = x, P = p, D = d] = x'\beta_d + K_d(p), \quad d \in \{0, 1\},$$

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 (D.11).*

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

(a) Estimate $\mathbb{E}[Y|P]$ nonparametrically,[9] and calculate the residual $e_i^P \equiv Y_i - \hat{\mathbb{E}}[Y_i|\hat{P}_i]$.

(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 estimates 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 $\widetilde{Y}$.

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

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).

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:

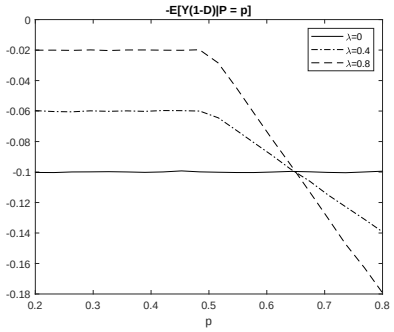
$$p_j = p_a + \frac{j-1}{J-1}(1 - p_a - p_n)$$

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 condition:

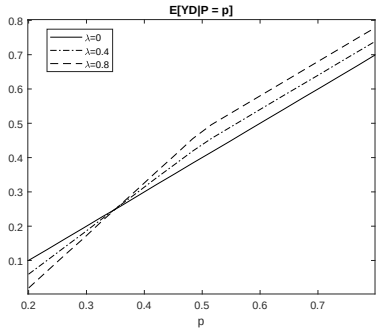
$$\mathbb{E}\left[Y \mid p_j\right] = \frac{1 - (1-\lambda)(p_n + p_a)}{1 - (p_n + p_a)} p_j - \frac{\lambda}{1 - (p_n + p_a)} p_a.$$

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 judge leniency design 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. Note that the parameter λ directly governs the shape of the function $\mathbb{E}[Y|P = p]$. The nonzero value of λ can potentially be generated by violations of one of the three assumptions (or

their combinations).



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



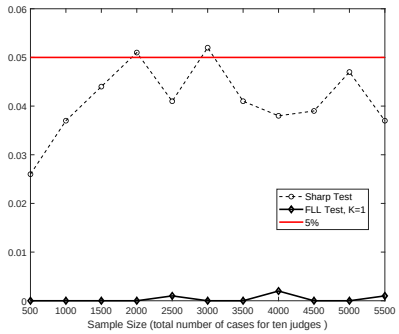
(b) $\mathbb{E}[YD|P=p]$

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

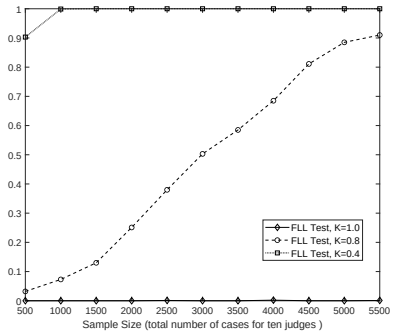
Figure 1 visually illustrates our testable implications of the judge leniency design for the specific function $g(Y) = 1\{0 < Y \leq 1\} = Y$ (because Y is binary). The left and right panels of the figure, respectively, depict $\mathbb{E}[-Y(1-D)|P=p]$ and $\mathbb{E}[YD|P=p]$. These population quantities 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 that both tests control size well in the aforementioned DGP. Specifically, it is evident from the graph that the rejection rate of our sharp test is controlled by and close to the nominal level of 5%. Conversely, the nonparametric test proposed in FLL consistently yields rejection rates close to zero when setting the tuning parameter $K = 1$.[10]

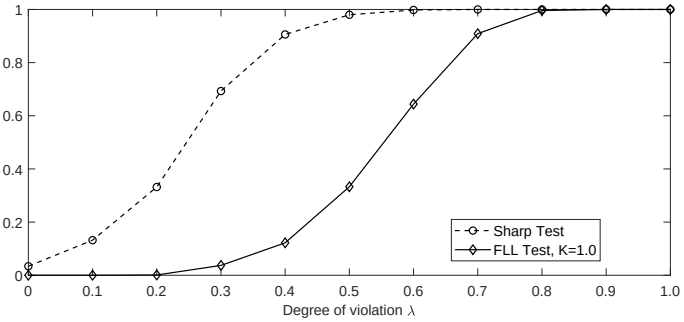
[10]Recall the outcome variable is binary; hence, the largest possible absolute value for the treatment



(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 two additional cases: when the maximum allowable treatment effect K is set at 0.8 and 0.4, respectively. The striking observation is that 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 re-

effect is 1. These results correspond to Figures 9 and 10 in the online appendix of FLL, where the rejection probabilities are nearly zero for various sample sizes when $\lambda = 0$.

striction 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, we set

$$Z = F_{Z^*}^{-1}\left(\frac{\ell(Z^*)}{L}\right), \quad \ell(Z^*) = \underset{\ell=1,2,\cdots,L-1}{\operatorname{argmin}} \left| F_{Z^*}(Z^*) - \frac{\ell}{L} \right|.$$

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

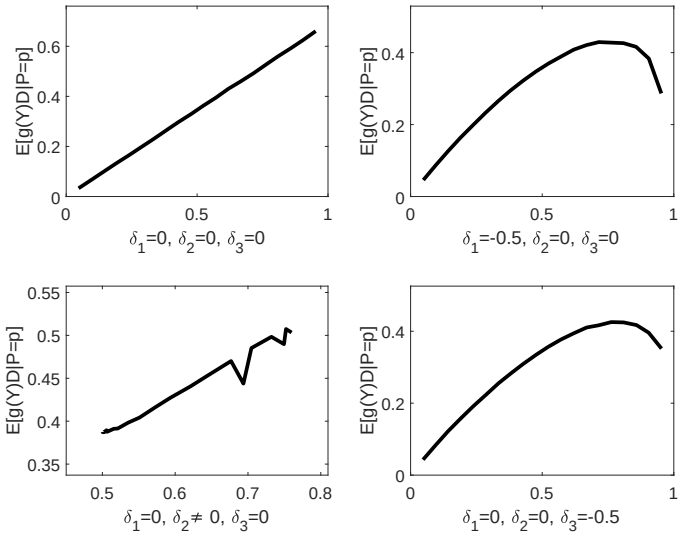


Figure 3: Sharp Testable Restrictions for Different DGPs

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, each with 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 8, 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) 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 that 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.

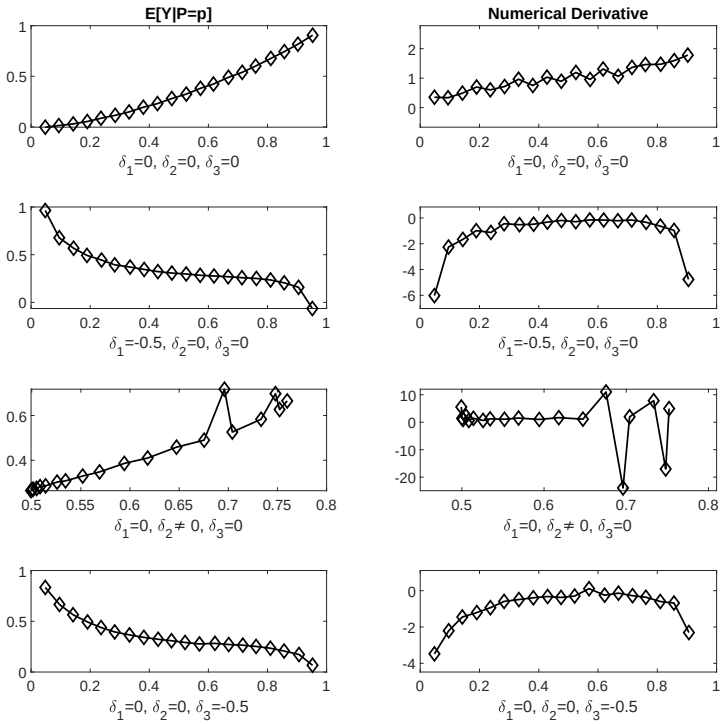


Figure 4: FLL Testable Restrictions for Different DGPs

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 ($\beta_1 = \beta_0 = 0$). When there are covariates, we use Carr and Kitagawa (2021)'s method to control for covariates, as discussed in the previous section. 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.

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

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 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 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.

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 video conference interview by Pretrial Services, which assesses risk factors and financial details for public defense eligibility. Utilizing this information, Pretrial Services assigns arrestees to a bail recommendation grid. Bail hearings,

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

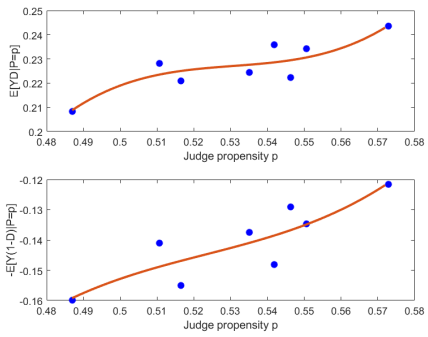
$\delta_2 = \delta_3 = 0$, $n = 1000$	$\delta_1 = -0.1$	$\delta_1 = -0.3$	$\delta_1 = -0.5$	$\delta_1 = -0.7$
Sharp Test	0.001	0.085	0.825	1.000
FLL-nonp, $K = \Delta_y$	0.000	0.000	0.000	0.000
FLL-nonp, $K = \frac{\Delta_y}{8}$	0.001	0.004	0.054	0.911
FLL-nonp, $K = \frac{\Delta_y}{16}$	0.026	0.006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

conducted by magistrates every four hours via video conference, involve a brief process where charges are explained, next court appearances are specified, eligibility for a court-appointed defense 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.

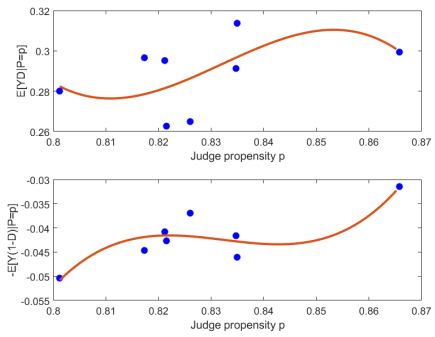
Stevenson (2018)'s research design leverages the varying magistrate tendencies to assign affordable bail as an instrument to study detention's impact on case outcomes. To answer the research questions, the author utilizes data from the court records of the Pennsylvania Unified Judicial System, obtained 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. As noted in Stevenson (2018), the shift-rotation system at the Philadelphia court forms the basis for such randomness.

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 two scatter plots for each crime category: $\{(p_j, \mathbb{E}[YD|P = p_j])\}_{j=1}^{8}$ and $\{(p_j, -E[Y(1-D)|P = p_j])\}_{j=1}^{8}$, along with a fitted

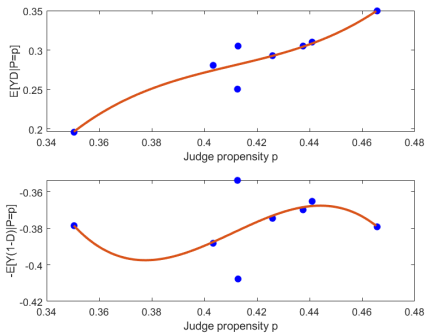
polynomial to illustrate whether the anticipated implications of the judge leniency design framework are satisfied for the considered categories of criminal cases. The graphs indicate $\mathbb{E}[YD|P = p]$ and $\mathbb{E}[-Y(1 - D)|P = p]$ are most likely to be non-decreasing for the aggressive assault case.[11] 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 that the assumptions are the least likely to be violated in the aggressive assault case, while the drug possession case shows the highest likelihood of violating the assumptions of the judge leniency design.



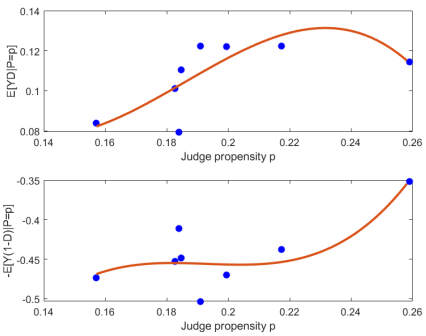
(a) Aggressive assault



(b) Robbery



(c) Drug sell



(d) Drug possession

Figure 5: Testable restrictions by case types

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

[11] 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.

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 offenses, our test rejects the validity of the judge leniency design. Meanwhile, FLL's semi-parametric test rejects the category of aggregate assault.[12] These results suggest that 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 dataset. 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 that controlling the crime type is important in practice.

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.

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

5 Salvage the Model under Weaker Assumptions

The rejection of the sharp test means that the judge's leniency design assumptions are too stringent for the data. In this case, relaxing some of these assumptions is required to salvage the model. There are different ways to relax a model's assumptions. One way is to maintain the same estimand used in the stringent model and ask under what conditions this estimand can still be interpreted causally. The model relaxation recently entertained by FLL falls into this second approach, providing alternative conditions under which the 2SLS could still have a causal interpretation when Assumptions 2.1 to 2.3 are too stringent for the data. In Section 5.1, we revisit the average exclusion assumption proposed by FLL and show it is a special case of a zero-covariance condition: a restriction that may not always be justifiable in all empirical settings.

There is another approach that focuses on a well-defined policy-relevant parameter and examines how this parameter could be point-identified or set-identified using weaker and more credible assumptions. In such a case, the parameter of interest remains the same, but the (set) estimands may vary depending on the credible assumptions one would be willing to maintain. We will discuss this approach in Section 5.2.[13]

5.1 Average Exclusion and Monotonicity

We first revisit the average exclusion and monotonicity conditions. For simplicity, suppose Z has finite support as in FLL such that $\mathcal{Z} = \{1, 2, \cdots, J\}$. The general form of the potential outcome model is,

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

FLL proposes to relax Assumption 2.2 and Assumption 2.3 with the average exclusion restriction and the average monotonicity assumption, respectively:

Assumption 5.1 *Let* $\lambda_z = \Pr(Z = z)$, $p_z = \mathbb{E}[D_z]$, $p = \sum_{z \in \mathcal{Z}} \lambda_z p_z$, $\bar{D} = \sum_{z \in \mathcal{Z}} \lambda_z D_z$, *and* $\bar{Y}_d = \sum_{z \in \mathcal{Z}} \lambda_z Y_d(z)$ *for* $d \in \{0, 1\}$.

[13] In this section, we mainly focus on the case in which the exclusion or monotonicity assumption is violated. When the random assignment assumption is violated, one can consider a partial identification, see Mourifié and Wan (2025).

(a) Average exclusion restriction:

$$\mathbb{E}\left[\sum_{z\in\mathcal{Z}}\lambda_z\left(p_z-p\right)\left\{(Y_0(z)-\bar{Y}_0)(1-D_z)+(Y_1(z)-\bar{Y}_1)D_z\right\}\right]=0.$$

(b) Average monotonicity: $\omega\equiv\sum_{z\in\mathcal{Z}}\lambda_z\left(p_z-p\right)\left(D_z-\bar{D}\right)\geq 0$ *almost surely.*

Under Assumptions 2.1 and 5.1, FLL's Theorem 3 shows that the 2SLS estimand (of using $P(Z)$ as IV) has a causal interpretation since it can be written as a weighted average of the following treatment effect $\delta=\bar{Y}_1-\bar{Y}_0$ i.e.,

$$\frac{\text{Cov}(Y,P(Z))}{\text{Cov}(D,P(Z))}=\mathbb{E}\left[\frac{\omega}{\mathbb{E}[\omega]}\delta\right]. \tag{5.1}$$

Note that the δ in Equation (5.1) is a deterministic function of the collection of potential outcomes $\{Y_d(z)\}_{d=0,1,z\in\mathcal{Z}}$. The 2SLS estimand is causal because it is a weighted average of δ and the weight ω is positive by the average monotonicity (Assumption 5.1-(b)).

Proposition 5.1 below provides a generalization and more transparent discussion of the FLL's Theorem 3. First, we demonstrate that the average exclusion assumption is essentially equivalent to a zero covariance condition. Second, we show that Equation (5.1) indeed holds for any deterministic function of $\{Y_d(z)\}_{d=0,1,z\in\mathcal{Z}}$, not just for δ.

To clarify these points, let us define $\alpha_z\equiv Y_1(z)-Y_0(z)$, and $\tilde{\alpha}=\sum_{z\in\mathcal{Z}}\alpha_z 1\{Z=z\}=Y_1(Z)-Y_0(Z)$. Let $\alpha\equiv h(Y_1(1),...,Y_1(J),Y_0(1),...,Y_0(J))$ be an arbitrary measurable deterministic function of the collection of potential outcomes. δ defined in Assumption 5.1 is a special case when we pick $h(Y_1(1),...,Y_1(J),Y_0(1),...,Y_0(J))=\bar{Y}_1-\bar{Y}_0$. One could instead be interested in different treatment effects specific to each judge: $\alpha=\alpha_z$, $z=1,2,\cdots,J$. α can also be a quantity without clear economic interpretation such as $\alpha=\sum_{z\in\mathcal{Z}}zY_1(z)$.

Proposition 5.1

(a) Under Assumption 2.1, and Assumption 5.1(b), the following equation holds for any measurable deterministic function α *of the collection of potential outcomes:*

$$\frac{Cov(Y,P(Z))}{Cov(D,P(Z))}=\mathbb{E}\left[\frac{\omega}{\mathbb{E}[\omega]}\alpha\right]+\frac{Cov\left((\tilde{\alpha}-\alpha)D+\tilde{Y}_0,P(Z)\right)}{\mathbb{E}\left[\omega\right]}$$

where $\tilde{Y}_d = \sum_{z \in \mathcal{Z}} Y_d(z) 1\{Z = z\}$.

(b) Under Assumption 2.1, for $\alpha = \bar{Y}_1 - \bar{Y}_0 = \sum_z \lambda_z (Y_1(z) - Y_0(z)) = \sum_z \lambda_z \alpha_z$ *we have:*

$$Cov\left((\tilde{\alpha} - \alpha)D + \tilde{Y}_0, P(Z)\right) = \mathbb{E}\left[\sum_{z \in \mathcal{Z}} \lambda_z \left(p_z - p\right) \left\{(Y_0(z) - \bar{Y}_0)(1 - D_z) + (Y_1(z) - \bar{Y}_1)D_z\right\}\right].$$

The proof for the proposition is collected in Appendix B.3. Under Assumption 2.1 (independence), Proposition 5.1(b) shows that the average exclusion restriction of FLL is, indeed, a special case of the zero-covariance assumption when $\alpha = \delta$. Proposition 5.1(a) further shows that if one targets an arbitrary quantity $\alpha \equiv h(Y_1(1), ..., Y_1(J), Y_0(1), ..., Y_0(J))$, and if one is willing to impose the same zero-covariance assumption on α:

$$\text{Cov}\left((\tilde{\alpha} - \alpha)D + \tilde{Y}_0, P(Z)\right) = 0, \tag{5.2}$$

then one can always interpret 2SLS estimand as the weighted average of α with positive weights under the average monotonicity Assumption 5.1(b). This happens because the above zero-covariance condition in Equation (5.2) is a reduced-form condition, which assumes that the correlation between a reduced-form error (involving the parameter of interest) and the propensity score, i.e. $\text{Cov}\left(Y - \alpha D, P(Z)\right) = 0$.

How does one assess the plausibility of the average exclusion condition? FLL provides a heuristic argument.[14] However, this argument could also be invoked by anyone who wants to impose that Equation (5.2) holds for other $\alpha \neq \delta$. Also, it is difficult to justify why $\text{Cov}\left((\tilde{\delta} - \delta)D + \tilde{Y}_0, P(Z)\right) = 0$ but $\text{Cov}\left((\tilde{\alpha} - \alpha)D + \tilde{Y}_0, P(Z)\right) \neq 0$ for other $\alpha \neq \delta$.

Furthermore, it is worth noting that the average exclusion restriction is not invariant to a relabelling of the treatment. In other terms, this assumption may hold if the researcher defines the treatment as D equals 1 if incarceration and 0 if not, while it may not hold if the researcher recodes the treatment as D equals 1 if no incarceration and 0 if incarceration. Indeed, after a relabelling, the zero-covariance in Equation (5.2) becomes: $\text{Cov}\left((\tilde{\alpha} - \alpha)D + \tilde{Y}_1, P(Z)\right) = 0$. It follows that the average exclusion assumption is invariant to a relabelling if and only if $\text{Cov}\left(\tilde{Y}_1, P(Z)\right) = \text{Cov}\left(\tilde{Y}_0, P(Z)\right)$.

[14] Frandsen, Lefgren, and Leslie (2023, page 19): "*Average exclusion can be probed by examining the correlation between judge-level treatment propensity and judge-level averages of alternative channels through which judges may affect outcomes if such channels are observed. Average exclusion may be more plausible if these correlations are near zero.*"

Despite all those discussed above, we do observe a direct way to assess the validity of Assumptions 2.1 and 5.1. In fact, under these assumptions, there is:

$$\left|\frac{\text{Cov}(Y, P(Z))}{\text{Cov}(D, P(Z))}\right| \leq \mathbb{E}\left[\left|\frac{\omega}{\mathbb{E}[\omega]}\delta\right|\right] \leq \mathbb{E}\left[|\delta|\right] \leq U - L$$

where U and L are, respectively, the upper and lower bounds of $\mathcal{Y}$. Therefore, if the support of the outcome is bounded from both above and below, then the absolute value of the 2SLS estimand must also be bounded.

5.2 Conditioning on Judge's Characteristics

In practice, it is not uncommon for researchers to have good reason to believe the assumptions hold after controlling for the judge's specific characteristics. We explore this idea in this section and demonstrate that it is closely related to the partial exclusion assumption (defined below) and the partial monotonicity assumption made in Mogstad, Torgovitsky, and Walters (2019). Specifically, we decompose Z into two components: Z_I and Z_c, and we assume the monotonicity and exclusion restriction hold conditionally on Z_c. Here, Z_c can be a judge's race or political party, and Z_I is a vector of the remaining characteristics.

Assumption 5.2 (Partial Exclusion) *Let* $Z \equiv (Z_I', Z_c')'$. *For* $d \in \{0, 1\}$, $Y_d(z) = Y_d(z_c)$ *for all* $z \in \mathcal{Z}$.

Assumption 5.3 (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.*

The partial exclusion assumption relaxes Assumption 2.2 and allows the potential outcomes to depend on the subvector Z_c. For instance, when the treatment of interest is incarceration, judges could assign and differ in other punishments, such as probation, fines, or sentence length. These other punishments could directly affect potential outcomes, making Assumption 2.2 unlikely. 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 future 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 states that whether and how the judge assigns other

types of punishment depends only on a subset of the judge's observable characteristics (Z_c), but not on others (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 characteristics $Z_c = z_c$. 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 partial monotonicity Assumption 5.3 was initially introduced in Mogstad, Torgovitsky, and Walters (2019). It 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 in the same order. Mathematically, we can not have both $\mathbb{P}(D(z_I, 1, 0) = 1, D(z_I', 0, 1) = 0) > 0$ and $\mathbb{P}(D(z_I', 0, 1) = 1, D(z_I, 1, 0) = 0) > 0$. However, there is a large body of empirical evidence of heterogeneity in the ranking of judges' leniency across different types of offense or defendants (see Abrams, Bertrand, and Mullainathan, 2012; Stevenson, 2018). This is, however, compatible with the partial monotonicity. Its main advantage is that it no longer requires a uniform ranking of defendants across different judges. Judges' rankings are allowed to vary with their characteristics Z_c. Applying the result of Vytlacil (2002), the partial monotonicity condition can be characterized as a partial single threshold-crossing restriction under the independence assumption Assumption 2.1, which we restated below.

Assumption 5.4 (Partial Single Threshold-Crossing) *Type $Z = (Z_I, Z_c)$'s judge treatment assignment mechanism is governed by the following threshold crossing model $D_z = 1\{\nu(Z_I, Z_c) \geq U_{Z_c}\}$ 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 5.4, we can apply the standard normalization,

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

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 face 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, 2024). 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 incarcerating a defendant. The particularity of the partial STC is that the expected cost can vary across judges with distinct observable characteristics z_c, but is fixed within judges with the same z_c. In the standard monotonicity assumption, the cost V would be the same regardless of the characteristics (z_I, z_c). For the same reason, the partial STC is also more reasonable in settings where decision-makers (judges) differ in their preferences and skills (Chan, Gentzkow, and Yu, 2022).

Here, we provide an example of eight judges deciding whether to incarcerate a given defendant to elucidate further 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 result in four types of judges. 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 the defendant in a separate unit segment for each judge. For example, p_{11} and p'_{11} are the benefits 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-colored, while those who release the defendant are red-colored.

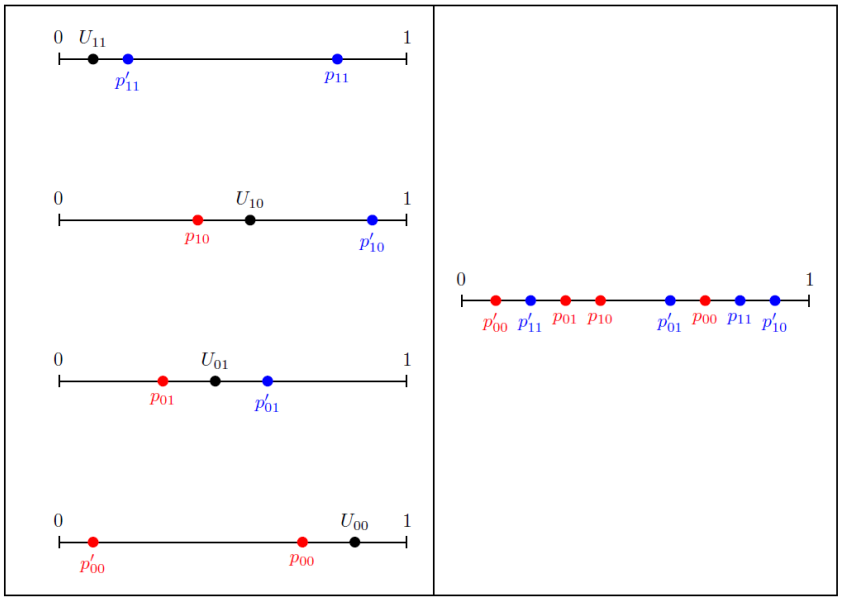


Figure 6: Monotonicity in Judge IV and Conditional Judge IV designs

The behavior of the eight judges does not violate Assumption 5.3 or Assumption 5.4. However, the standard monotonicity Assumption 2.3 is clearly violated (right rectangle of Figure 6). Indeed, the judge with propensity score p'_{11} incarcerates the defendant (blue-colored), whereas judges with higher propensity scores p_{01}, p_{10}, or p_{00} do not incarcerate the defendant (red-colored). 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 5.3 (or Assumption 5.4) 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, 5.2 and 5.4 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 under Partial exclusion and monotonicity) *If Assumptions 2.1, 5.2 and 5.4 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 $(p, p') \in \mathcal{P}_{z_c} \times \mathcal{P}_{z_c}$ such that $p < p'$ we have the following identification results:

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

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

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

(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 that 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 that 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 that 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 highlight the importance of considering sharp implications for a better use of information in the data. While we focus on the primary application of testing the validity of judge leniency design, our method can be readily applied to a broad range of other applications.

APPENDIX

A Implementation of the test

In this section, we describe the details of calculating the test statistics for Algorithm 3.1. Let $\{Y_i, Z_i, D_i\}_{i=1}^n$ be a random sample and $P(D_i = 1|Z_i) = P(Z_i, \theta_0)$ be the propensity known to θ_0. Note that when Z is the judge's identity and if the number of defendants for each judge diverges to infinity, we can simply use the frequency estimator $\hat{P}_i = \frac{\sum_{k=1}^n D_k 1\{Z_k=Z_i\}}{\sum_{k=1}^n 1\{Z_k=Z_i\}}$. Therefore, in the appendix sections, we focus on the case in which Z is continuous to simplify notation.

A.1 Constructing $\hat{\nu}_d(\ell)$

First, when Z is continuous, we estimate θ_0 by MLE,

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

where $P(z, \theta)$ is parameterized and depends on z through $z'\theta$. For example, $P(z, \theta) = \Phi(z'\theta_z)$ for Probit or $P(z, \theta) = \frac{exp(z'\theta_z)}{1+\exp(z'\theta_z)}$ for Logit.

Next, note that

$$\begin{aligned}\nu_1(y, r_y, p_1, p_2, r_p, \theta_0) &= m_1(y, r_y, p_2, r_p, \theta_0) \cdot w(p_1, r_p, \theta_0) - m_1(y, r_y, p_1, r_p, \theta_0) \cdot w(p_2, r_p, \theta_0), \\ \nu_0(y, r_y, p_1, p_2, r_p, \theta_0) &= m_0(y, r_y, p_2, r_p, \theta_0) \cdot w(p_1, r_p, \theta_0) - m_0(y, r_y, p_1, r_p, \theta_0) \cdot w(p_2, r_p, \theta_0),\end{aligned}$$

where

$$m_1(y, r_y, p, r_p, \theta) = \mathbb{E}[D1(y \le Y \le y + r_y)1(p \le P(Z, \theta) \le p + r_p)], \tag{A.2}$$

$$m_0(y, r_y, p, r_p, \theta) = \mathbb{E}[(D - 1)1(y \le Y \le y + r_y)1(p \le P(Z, \theta) \le p + r_p)], \tag{A.3}$$

$$w(p, r_p, \theta) = \mathbb{E}[1(p \le P(Z, \theta) \le p + r_p)]. \tag{A.4}$$

We can estimate $m_d(y, r_y, p, r_p, \theta)$ and $w(p, r_p, \theta)$ by sample analogs and θ be replaced by its

MLE $\hat{\theta}$:

$$\hat{m}_d(y, r_y, p, r_p, \hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n} m_{di}(y, r_y, p, r_p, \hat{\theta}), \quad d = 0, 1 \tag{A.5}$$

$$\hat{w}(p, r_p, \hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n} w_i(p, r_p, \hat{\theta}). \tag{A.6}$$

with

$$\begin{aligned}
m_{1i}(y, r_y, p, r_p, \theta) &= D_i 1(y \le Y_i \le y + r_y) 1(p \le P(Z_i, \theta) \le p + r_p), \\
m_{0i}(y, r_y, p, r_p, \theta) &= (D_i - 1) 1(y \le Y_i \le y + r_y) 1(p \le P(Z_i, \theta) \le p + r_p), \\
w_i(p, r_p, \theta) &= 1(p \le P(Z_i, \theta) \le p + r_p).
\end{aligned}$$

Then, for a given $\ell = (y, r_y, p_1, p_2, r_p)'$, we can estimate $\nu_1(\ell)$ and $\nu_0(\ell)$ by

$$\hat{\nu}_1(\ell) = \hat{m}_1(y, r_y, p_2, r_p, \hat{\theta}) \cdot \hat{w}(p_1, r_p, \hat{\theta}) - \hat{m}_1(y, r_y, p_1, r_p, \hat{\theta}) \cdot \hat{w}(p_2, r_p, \hat{\theta}), \tag{A.7}$$

$$\hat{\nu}_0(\ell) = \hat{m}_0(y, r_y, p_2, r_p, \hat{\theta}) \cdot \hat{w}(p_1, r_p, \hat{\theta}) - \hat{m}_0(y, r_y, p_1, r_p, \hat{\theta}) \cdot \hat{w}(p_2, r_p, \hat{\theta}). \tag{A.8}$$

A.2 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 both mean and variance equal to one. For instance, we can choose standard normal. Let $\hat{\theta}^b$ be the MLE based on the b-th bootstrapped sample:

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

and the estimated propensity score for the b-th bootstrap as

$$\hat{P}_i^b = P(Z_i, \hat{\theta}^b) \tag{A.9}$$

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

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

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

$$\hat{\nu}_1^b(\ell) = \hat{m}_1^b(y, r_y, p_2, r_p, \hat{\theta}) \cdot \hat{w}^b(p_1, r_p, \hat{\theta}^b) - \hat{m}_1^b(y, r_y, p_1, r_p, \hat{\theta}^b) \cdot \hat{w}^b(p_2, r_p, \hat{\theta}^b), \tag{A.10}$$
$$\hat{\nu}_0^b(\ell) = \hat{m}_0^b(y, r_y, p_2, r_p, \hat{\theta}^b) \cdot \hat{w}^b(p_1, r_p, \hat{\theta}^b) - \hat{m}_0^b(y, r_y, p_1, r_p, \hat{\theta}^b) \cdot \hat{w}^b(p_2, r_p, \hat{\theta}^b). \tag{A.11}$$

B Proof of Main Results

B.1 Proof of Theorem 1

Proof. Theorem 1-(i) is a direct application of Heckman and Vytlacil (2005)'s testable implications where $g(Y) = 1\{Y \in (y, y']\}$ for $y \leq y'$. We focus on part (ii).

We define some notation. Let $\mathcal{L}(\mathcal{P})$ be the set of limit points of $\mathcal{P}$, $\mathcal{L}^o(\mathcal{P})$ be a set of interior point of $\mathcal{P}$, and $\mathcal{C}(\mathcal{P})$ be the closure of $\mathcal{P}$. Furthermore, let $I(\mathcal{P}) = \mathcal{C}(\mathcal{P})/\mathcal{L}^o(\mathcal{P})$ be the complement of $\mathcal{L}^o(\mathcal{P})$ in the closure of $\mathcal{P}$. So $I(\mathcal{P})$ also contains isolation points. Note that $\mathcal{L}^o(\mathcal{P})$ can be written as a union of countable or finite exclusive open intervals: $\mathcal{L}^o(\mathcal{P}) = \cup_{j=1}^J (a_j, b_j)$, where $(a_j, b_j) \subseteq \mathcal{P}$, $b_j < a_{j+1}$, and J can be infinity. Let $\Omega(\mathcal{P})$ be a collection of intervals belonging to $(0, 1]$ defined as follows:

$$\Omega(\mathcal{P}) \equiv \big\{(p, p'] : p, p' \in I(\mathcal{P}) \cup \{0, 1\} \text{ and } (p, p') \cap \mathcal{P} = \emptyset\big\}.$$

So the interior of each interval does not intersect with $\mathcal{P}$. $\Omega(\mathcal{P})$ contains a generic element $(c_k, d_k]$, where c_k, $d_k \in I(\mathcal{P})$, $d_k \leq c_{k+1}$, $k = 1, 2, \cdots, K$ with K possibly equals to ∞, depending on how many isolation points there are in $\mathcal{P}$. Note that with above notation, for any $v \in (0, 1]$, v must belongs to one of the following categories: (i) an element of $\mathcal{L}^o(\mathcal{P})$ so that $v \in (a_j, b_j)$ for some j, (ii) $v \in \mathcal{L}(\mathcal{P})/\mathcal{L}^o(\mathcal{P})$, and (iii) there exist an integer k such that $v \in (c_k, d_k]$. The following figure illustrates the partition of the unit interval.

Figure 7: An illustration: $\mathcal{P} = \{p_1, p_2, p_5\} \cup [p_3, p_4] \cup [p_6, p_7]$, $\mathcal{L}^o(\mathcal{P}) = (p_3, p_4) \cup (p_6, p_7)$, and $\Omega(\mathcal{P}) = \{(0, p_1], (p_1, p_2], (p_4, p_5], (p_5, p_6], (p_7, 1]\}$.

We will assume that $\mathbb{P}(y < Y \leq y', D = 1|P = p)$ and $\mathbb{P}(y < Y \leq y', D = 0|P = p)$ are continuously differentiable over $\mathcal{L}^o$ as a regularity condition under which the local instrumental variable (LIV) estimand is well defined.

First, we construct $\tilde{V}$ and $\tilde{D}$ as follows:

$$\mathbb{P}(\tilde{V} \leq t|P = p) = t, \forall (t, p) \in [0, 1] \times \mathcal{P}, \text{ and } \tilde{D} = 1\{P(Z) \geq \tilde{V}\}.$$

By construction, Assumption 2.4 is satisfied. Next, we propose the following distribution for $\tilde{Y}_1|\tilde{V}, P$. For any arbitrary $p \in \mathcal{P}$ and $v \in (0, 1]$, we define

$$\mathbb{P}(\tilde{Y}_1 \leq y|\tilde{V} = v, P = p) = \begin{cases} \frac{\partial}{\partial t}\mathbb{P}(Y \leq y, D = 1|P = t)|_{t=v} & \text{if } v \in \mathcal{L}^o(\mathcal{P}) \\ \lim_{\tilde{v} \to v} \frac{\partial}{\partial t}\mathbb{P}(Y \leq y, D = 1|P = t)|_{t=\tilde{v}} & \text{if } v \in \mathcal{L}(\mathcal{P})/\mathcal{L}^o(\mathcal{P}) \\ \frac{\mathbb{P}(Y \leq y, D=1|P=d_k) - \mathbb{P}(Y \leq y, D=1|P=c_k)}{d_k - c_k} & \text{if } v \notin L(P) \text{ but } v \in (c_k, d_k] \in \Omega(\mathcal{P}). \end{cases}$$

$$\mathbb{P}(\tilde{Y}_0 \leq y|\tilde{V} = v, P = p) = \begin{cases} -\frac{\partial}{\partial v}\mathbb{P}(Y \leq y, D = 0|P = t)|_{t=v} & \text{if } v \in \mathcal{L}^o(\mathcal{P}) \\ -\lim_{\tilde{v} \to v} \frac{\partial}{\partial v}\mathbb{P}(Y \leq y, D = 0|P = t)|_{t=\tilde{v}} & \text{if } v \in \mathcal{L}^o(\mathcal{P}) \\ \frac{\mathbb{P}(Y \leq y, D=0|P=c_k) - \mathbb{P}(Y \leq y, D=0|P=d_k)}{d_k - c_k} & \text{if } v \notin L^o(P) \text{ but } v \in (c_k, d_k] \in \Omega(\mathcal{P}). \end{cases}$$

Note that the conditioning on $\tilde{V} = v$ and $P = p$, the distribution of $\tilde{Y}_1$ does not depend on p. Hence, Assumption 2.1 is satisfied by construction.

We now show that the distribution function constructed above is well defined. We focus on $\mathbb{P}(\tilde{Y}_1 \leq y|\tilde{V} = v, P = p)$ and the verification for $\mathbb{P}(\tilde{Y}_0 \leq y|\tilde{V} = v, P = p)$ is analogous. Let $\underline{y}$ and $\overline{y}$ be the lower and upper bounds of the support of Y, respectively.

1. $\mathbb{P}(\tilde{Y}_1 < \underline{y} - \epsilon|\tilde{V} = v, P = p) = 0$ for all $v \in [0, 1]$ and for any arbitrarily small $\epsilon > 0$. To see

this, suppose $v \notin \mathcal{L}(\mathcal{P})$, then there exists $(c_k, d_k] \in \Omega(\mathcal{P})$ such that $v \in (c_k, d_k]$, therefore,

$$\begin{aligned}&\mathbb{P}(\tilde{Y}_1 \leq \underline{y} - \epsilon | \tilde{V} = v, P = p) \\ =&\frac{\mathbb{P}(Y \leq \underline{y} - \epsilon, D = 1 | P = d_k) - \mathbb{P}(Y \leq \underline{y} - \epsilon, D = 1 | P = c_k)}{d_k - c_k} = \frac{0 - 0}{d_k - c_k} = 0.\end{aligned}$$

On the other hand, if $v \in \mathcal{L}^o(\mathcal{P})$, then $\mathbb{P}(Y \leq \underline{y} - \epsilon, D = 1 | P = \tilde{v}) = 0$ for all $\tilde{v}$ in a small neighborhood of v, which implies $\frac{\partial}{\partial v}\mathbb{P}(Y \leq \underline{y} - \epsilon, D = 1 | P = v) = 0$. The case that $v \in \mathcal{L}^o(\mathcal{P})$ follows straightforwardly.

2. $\mathbb{P}(\tilde{Y}_1 \leq \overline{y} | \tilde{V} = v, P = p) = 1$. First, if $v \in \mathcal{L}^o(\mathcal{P})$, then

$$\mathbb{P}(Y \leq \overline{y}, D = 1 | P = v) = \mathbb{P}(D = 1 | P = v) = v \Rightarrow \frac{\partial}{\partial v}\mathbb{P}(Y \leq \overline{y}, D = 1 | P = v) = 1.$$

On the other hand, if $v \notin \mathcal{L}(\mathcal{P})$, then

$$\mathbb{P}(\tilde{Y}_1 \leq \overline{y} | \tilde{V} = v, P = p) = \frac{\mathbb{P}(Y \leq \overline{y}, D = 1 | P = d_k) - \mathbb{P}(Y \leq \overline{y}, D = 1 | P = c_k)}{p' - p} = \frac{d_k - c_k}{d_k - c_k} = 1.$$

3. $\mathbb{P}(\tilde{Y}_1 \leq y | \tilde{V} = v, P = p)$ is nondecreasing in y. For $y < y'$ we have

$$\begin{aligned}&\mathbb{P}(\tilde{Y}_1 \leq y' | \tilde{V} = v, P = p) - \mathbb{P}(\tilde{Y}_1 \leq y | \tilde{V} = v, P = p) \\ =&\begin{cases} \frac{\partial}{\partial t}\mathbb{P}(y < Y \leq y', D = 1 | P = t)|_{t=v} \geq 0 \quad \text{if } v \in \mathcal{L}^o(\mathcal{P}), \\ \lim_{\tilde{v} \to v} \frac{\partial}{\partial t}\mathbb{P}(y < Y \leq y, D = 1 | P = t)|_{t=\tilde{v}} \geq 0 \quad \text{if } v \in \mathcal{L}(\mathcal{P}) / \mathcal{L}^o(\mathcal{P}) \\ \frac{\mathbb{P}(y < Y \leq y', D = 1 | P = d_k) - \mathbb{P}(y < Y \leq y', D = 1 | P = c_k)}{d_k - c_k} \geq 0 \text{ if } v \notin L^o(P) \text{ but } v \in [c_k, d_k] \in \Omega(\mathcal{P}), \end{cases}\end{aligned}$$

where the last inequalities hold whenever the testable implications hold, i.e. $\mathbb{P}(y < Y \leq y', D = 1 | P = p)$ is a non-decreasing function for all $p \in \mathcal{P}$ and all $y < y'$, and by the continuous differentiability of $\mathbb{P}(y < Y \leq y', D = 1 | P = p)$ over $\mathcal{L}(\mathcal{P})$.

Finally, we show that $(\tilde{V}, \tilde{Y}_d, P(Z))$, $d \in \{0, 1\}$ is observationally equivalent to $(V, Y_d, P(Z))$ $d \in \{0, 1\}$. For this, we show that the conditioning distribution of $(\tilde{Y}, \tilde{D})$ given $P(Z)$ is the same as the conditioning of (Y, D) given $P(Z)$. Take an arbitrary $p \in \mathcal{P}$.

Suppose first $p \notin \mathcal{L}^o(\mathcal{P})$, then $(0, p]$ can be expressed as unions of exclusive intervals $\left(\cup_{j=1}^{J^*}(a_j, b_j)\right) \cup \left(\cup_{k=1}^{K^*}(c_k, d_k]\right)$ for some J^* and K^*, where $(a_j, b_j)s$ are connected subsets of

$\mathcal{P}$. Therefore,

$$\begin{aligned}
\mathbb{P}(\tilde{Y} \le y, \tilde{D} = 1|P = p) &= \mathbb{P}(\tilde{Y}_1 \le y, \tilde{V} \le p|P = p) = \int_0^p \mathbb{P}(\tilde{Y}_1 \le y|\tilde{V} = v, P = p)dv \\
&= \sum_{j=1}^{J^*} \int_{a_j}^{b_j} \mathbb{P}(\tilde{Y}_1 \le y|\tilde{V} = v, P = p)dv + \sum_{k=1}^{K^*} \int_{c_k}^{d_k} \mathbb{P}(\tilde{Y}_1 \le y|\tilde{V} = v, P = p)dv \\
&= \sum_{j=1}^{J^*} (\mathbb{P}(Y \le y, D = 1|P = b_j) - \mathbb{P}(Y \le y, D = 1|P = a_j)) \\
&\quad + \sum_{k=1}^{K^*} (\mathbb{P}(Y \le y, D = 1|P = d_k) - \mathbb{P}(Y \le y, D = 1|P = c_k)) \\
&= \mathbb{P}(Y \le y, D = 1|P = p) - \mathbb{P}(Y \le y, D = 1|P = 0) = \mathbb{P}(Y \le y, D = 1|P = p),
\end{aligned}$$

where the first equality is by construction that $\tilde{V}$ satisfies Assumption 2.4, the third equality holds because $(0, p]$ can be expressed as unions of exclusive intervals $\left(\cup_{j=1}^{J^*}(a_j, b_j)\right) \cup \left(\cup_{k=1}^{K^*}(c_k, d_k]\right)$, the fourth equality is obtained by inserting the constructed counterfactural distributions, and the last one holds because $\mathbb{P}(Y \le y, D = 1|P = 0) = 0$.

Suppose that $p \in (a_{j^*}, b_{j^*}) \subseteq \mathcal{L}^0(\mathcal{P})$ for some j^*, then the right hand side equals to

$$\begin{aligned}
\mathbb{P}(\tilde{Y} \le y, \tilde{D} = 1|P = p) &= \mathbb{P}(\tilde{Y}_1 \le y, \tilde{V} \le p|P = p) = \int_0^p \mathbb{P}(\tilde{Y}_1 \le y|\tilde{V} = v, P = p)dv \\
&= \int_0^{a_{j^*}} \mathbb{P}(\tilde{Y}_1 \le y|\tilde{V} = v, P = p)dv + \int_{a_{j^*}}^p \mathbb{P}(\tilde{Y}_1 \le y|\tilde{V} = v, P = p)dv \\
&= \mathbb{P}(Y \le y, D = 1|P = a_{j^*}) + \int_{a_{j^*}}^p \frac{\partial}{\partial v}\mathbb{P}(Y \le y, D = 1|P = v)dv \\
&= \mathbb{P}(Y \le y, D = 1|P = a_{j^*}) + \mathbb{P}(Y \le y, D = 1|P = p) - \mathbb{P}(Y \le y, D = 1|P = a_{j^*}) \\
&= \mathbb{P}(Y \le y, D = 1|P = p),
\end{aligned}$$

where the $\int_0^{a_{j^*}} \mathbb{P}(\tilde{Y}_1 \le y|\tilde{V} = v, P = p)dv = \mathbb{P}(Y \le y, D = 1|P = a_{j^*})$ holds by the above argument and the fifth equality holds by inserting the constructed counterfactural distributions. This completes the proof.□

B.2 Proof of Theorem 2

We begin by listing a few regularity conditions for the proof of Theorem 2. Again, when Z is the judge's identity, we use the frequency estimator $\hat{P}_i = \frac{\sum_{k=1}^n D_k 1\{Z_k = Z_i\}}{\sum_{k=1}^n 1\{Z_k = Z_i\}}$ for the propensity score. For its root-n-consistency, we only need $\sum_{i=1}^n 1\{Z_k = j\} \to \infty$ for each judge j, and i.i.d. of

(Y_i, D_i, X_i) among defendants conditioning on judges. So Assumptions B.1 to B.3 and B.5 are mostly for the case of continuous instrument Z.

Assumption B.1 *The observations $\{(Y_i, D_i, Z_i, X_i)\}_{i=1}^n$ are i.i.d. across i.*

For notational simplicity, Assumption B.1 assumes that all cases are mutually independent, which is equivalent to assuming that each judge handles exactly one case. All inference results can be extended straightforwardly to settings where judges handle multiple cases (with varying case counts across judges) by accounting for clustering at the judge level.

Assumption B.2 *We impose the following smoothness conditions:*

1. *The conditional density of (Y, D) given $P(Z, \theta_0) = p$, denoted by $f_{Y,D|P}(y, d|p)$, is Lipschitz continuous both in p on $\mathcal{P}$ and in y on $\mathcal{Y}$ for $d = 0, 1$.*

2. *For all $z \in \mathcal{Z}$, $P(z, \theta)$ is continuously differentiable in θ at θ_0 with bounded derivatives.*

Note that Assumption B.2-(1) does not exclude the case of discrete propensity score. When P is discrete and $\mathcal{P}$ contains finite many distinguished elements, any convergent sequence in $\mathcal{P}$ must be a constant sequence eventually, and in that case Assumption B.2-(1) holds automatically. Assumption B.2-(1) implies that the functions m_d and ω, defined in Equations (A.2) to (A.4), are continuous functions of ℓ. Assumption B.2-(2) implies that the class of functions $\{1(p \leq P(Z, \theta) \leq p + r_p) : \theta \in \Theta, p \in [0, 1], r_p \in [0, 1]\}$ is a Vapnik-Chervonenkis (VC) class of function.

Assumption B.3 *The parameter space Θ for θ_0 is compact, and θ_0 is in the interior of Θ. The estimator $\hat{\theta}$ admits an influence function of the following form,*

$$\sqrt{n}(\hat{\theta} - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} s(D_i, Z_i, \theta_0) + o_p(1), \tag{B.1}$$

where $s(\cdot, \cdot, \cdot)$ is measurable, satisfying $\mathbb{E}[s(D_i, Z_i, \theta_0)] = 0$, $\mathbb{E}[\sup_\theta |s(D_i, Z_i, \theta)|] < \infty$, and $V(\sup_\theta |s(D_i, Z_i, \theta)|) < \infty$.

Assumption B.3 is satisfied for common maximum likelihood estimators and parametric binary response models. For example, if one estimates θ_0 by Probit model $D_i = 1[Z_i'\theta_0 \geq V_i]$, with $V_i \sim N(0, 1)$, then

$$s(D_i, Z_i, \theta_0) = \frac{\phi((2D_i - 1)Z_i'\theta_0)}{\Phi((2D_i - 1)Z_i'\theta_0)} Z_i.$$

If the Logit model is used, then

$$s(D_i, Z_i, \theta_0) = \left(D_i - \frac{\exp(Z_i'\theta_0)}{1 + \exp(Z_i'\theta_0)} \right) Z_i.$$

Assumption B.4 *$\{W_i\}_{i=1}^n$ is a sequence of i.i.d. pseudo random variables that is independent of the sample path with $E[W_i] = 1$ and $Var[W_i] = 1$.*

Assumption B.5 *The estimator $\hat{\theta}^b$ satisfies that*

$$\sqrt{n}(\hat{\theta}^b - \hat{\theta}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (W_i - 1) \cdot s(D_i, Z_i, \theta_0) + o_p(1), \tag{B.2}$$

where $s_\theta(\cdot)$ is the same as in Assumption B.3.

Assumption B.5 is satisfied under our weighted bootstrap procedure.

The proof of Theorem Theorem 2 follows from the same arguments as Theorems 5.1 and 5.2 of Hsu (2017) once Lemmas D.1 to D.4 are established, as detailed in Appendix D.1.

B.3 Proofs for Proposition 5.1

Part (a). Note that,

$$\begin{aligned}
\text{Cov}(Y, P(Z)) =& \text{Cov}\left(\tilde{Y}_1 D + \tilde{Y}_0(1 - D), P(Z)\right) \\
=& \text{Cov}\left(\tilde{\alpha} D + \tilde{Y}_0, P(Z)\right) \\
=& \text{Cov}\left(\alpha D + (\tilde{\alpha} - \alpha) D + \tilde{Y}_0, P(Z)\right) \\
=& \text{Cov}(\alpha D, P(Z)) + \text{Cov}\left((\tilde{\alpha} - \alpha) D + \tilde{Y}_0, P(Z)\right).
\end{aligned}$$

The first term on the right hand side can be written as

$$\begin{aligned}
\text{Cov}(\alpha D, P(Z)) =& \mathbb{E}\left[\alpha D (P(Z) - p)\right] \\
=& \sum_{z=1}^{J} \mathbb{E}\left[\alpha D_z (p_z - p) \mid Z = z\right] \lambda_z \\
=& \mathbb{E}\left[\alpha \sum_{z=1}^{J} D_z (p_z - p) \lambda_z\right] = \mathbb{E}\left[\alpha \omega\right],
\end{aligned}$$

where the conditioning variable $Z = z$ is removed by the independence Assumption 2.1. This shows that

$$\text{Cov}(Y, P(Z)) = \mathbb{E}\left[\alpha \omega\right] + \text{Cov}\left((\tilde{\alpha} - \alpha) D + \tilde{Y}_0, P(Z)\right).$$

Next, it is easy to verify that

$$\text{Cov}(D, P(Z)) = \mathbb{E}\left[D(P(Z) - p)\right] = \mathbb{E}\left[\omega\right].$$

Finally, the proof is completed by taking the ratio of $\text{Cov}(Y, P(Z))$ and $\text{Cov}(D, P(Z))$ (which is possible as long as the instrument is relevant).

Part (b).

$$\begin{aligned}
\text{Cov}\left((\tilde{\alpha}-\alpha)D+\tilde{Y}_0, P(Z)\right) =& \mathbb{E}\left[\left((\tilde{\alpha}-\alpha)D+\tilde{Y}_0\right)(P(Z)-p)\right] \\
=& \sum_{z=1}^{J} \mathbb{E}\left[((\alpha_z-\alpha)D_z+Y_{0z})(p_z-p) \mid Z=z\right]\lambda_z \\
=& \sum_{z=1}^{J} \mathbb{E}\left[\lambda_z(p_z-p)((\alpha_z-\alpha)D_z+Y_{0z})\right] \\
=& \sum_{z=1}^{J} \mathbb{E}\left[\lambda_z(p_z-p)(Y_{1z}D_z+Y_{0z}(1-D_z)-\alpha D_z)\right] \\
=& \sum_{z=1}^{J} \mathbb{E}\left[\lambda_z(p_z-p)\left(Y_{1z}D_z+Y_{0z}(1-D_z)-\left(\bar{Y}_1 D_z+\bar{Y}_0(1-D_z)-\bar{Y}_0\right)\right)\right] \\
=& \sum_{z=1}^{J} \mathbb{E}\left[\lambda_z(p_z-p)\left((Y_{1z}-\bar{Y}_1)D_z+(Y_{0z}-\bar{Y}_0)(1-D_z)+\bar{Y}_0\right)\right] \\
=& \mathbb{E}\left[\sum_{z=1}^{J}\lambda_z(p_z-p)\left((Y_{1z}-\bar{Y}_1)D_z+(Y_{0z}-\bar{Y}_0)(1-D_z)\right)\right],
\end{aligned}$$

where the third equality is by the independenceAssumption 2.1, the fifth is by substituting for $\alpha = \bar{Y}_1 - \bar{Y}_0$, and the last equality holds because $\mathbb{E}\left[\sum_{z=1}^{J}\lambda_z(p_z-p)\bar{Y}_0\right] = 0$. □

B.4 Detailed derivation for Example 2.1

For the ease of reading, we restate the DGP below. Consider the potential outcome model:

$$\begin{cases} Y = Y_1 D + Y_0(1-D), \\ D = 1\{P \geq V\}. \end{cases}$$

We assume V is independent of (Y_1, Y_0, P). However, (Y_1, Y_0) and P are dependent:

$$\begin{cases} Y_1|P=\tilde{p} \sim \text{ degenerate at } 1\ , & \text{if } \tilde{p} < \frac{1}{2} \\ Y_1|P=\tilde{p} \sim Bernoulli(\tilde{p}), & \text{if } \tilde{p} \geq \frac{1}{2} \end{cases}$$

$$\begin{cases} Y_0|P=\tilde{p} \sim \text{ degenerate at } 0\ , & \text{if } \tilde{p} < \frac{1}{2} \\ Y_0|P=\tilde{p} \sim Bernoulli(\tilde{p}), & \text{if } \tilde{p} \geq \frac{1}{2} \end{cases}$$

We first check inequality (2.3). Let $p' > p > \frac{1}{2}$, and use the condition that V is independent of (Y_1, Y_0, P), we have

$$W(g(YD), p, p') = \frac{\mathbb{E}[YD|P = p'] - \mathbb{E}[YD|P = p]}{p' - p}$$
$$= \frac{\mathbb{E}[Y_1|P = p']p' - \mathbb{E}[Y_1|P = p]p}{p' - p} = \frac{p'^2 - p^2}{p' - p} = p' + p > 1 \equiv U_g.$$

Therefore, condition (2.3) is violated.

Next, we check condition (2.2). Based on the relative positoin of p', p, and $\frac{1}{2}$, we verify it by four cases.

(i) Suppose first $p' > \frac{1}{2} > p$,

$$W(g(Y), p, p') = \frac{\mathbb{E}[Y_1 D + (1 - D)Y_0|P = p'] - \mathbb{E}[Y_1 D + (1 - D)Y_0|P = p]}{p' - p}$$
$$= \frac{\mathbb{E}[Y_1|P = p']p' - \mathbb{E}[Y_1|P = p]p + \mathbb{E}[Y_0|P = p'](1 - p') - \mathbb{E}[Y_0|P = p](1 - p)}{p' - p}$$
$$= \frac{p'^2 - p + \mathbb{E}[Y_0|P = p'](1 - p') - \mathbb{E}[Y_0|P = p](1 - p)}{p' - p}$$
$$= \frac{p'^2 - p + p'(1 - p')}{p' - p} = 1 = U_g - L_g,$$

where $\mathbb{E}[Y_0|P = p] = 0$ because Y_0 is degenerate at 0 when conditioning on $P = p < \frac{1}{2}$, and $\mathbb{E}[Y_0|P = p'] = p'$ because $Y_0 \sim Bernoulli(p')$ when conditioning on $P = p' \geq \frac{1}{2}$.

(ii) Suppose $p > \frac{1}{2} > p'$,

$$W(g(Y), p, p') = \frac{\mathbb{E}[Y_1|P = p']p' - \mathbb{E}[Y_1|P = p]p + \mathbb{E}[Y_0|P = p'](1 - p') - \mathbb{E}[Y_0|P = p](1 - p)}{p' - p}$$
$$= \frac{p' - p^2 + \mathbb{E}[Y_0|P = p'](1 - p') - \mathbb{E}[Y_0|P = p](1 - p)}{p' - p}$$
$$= \frac{p' - p^2 - p(1 - p)}{p' - p} = 1 = U_g - L_g.$$

(iii) If $\frac{1}{2} > p' > p$, then

$$W(g(Y), p, p') = \frac{\mathbb{E}[Y_1|P = p']p' - \mathbb{E}[Y_1|P = p]p + \mathbb{E}[Y_0|P = p'](1 - p') - \mathbb{E}[Y_0|P = p](1 - p)}{p' - p}$$
$$= \frac{p' - p}{p' - p} = 1 = U_g - L_g,$$

because in this case Y_1 and Y_0 are degenerate at 1 and 0, respectively.

(iv) If $p' > p > \frac{1}{2}$, then

$$W(g(Y), p, p') = \frac{\mathbb{E}[Y_1|P = p']p' - \mathbb{E}[Y_1|P = p]p + \mathbb{E}[Y_0|P = p'](1 - p') - \mathbb{E}[Y_0|P = p](1 - p)}{p' - p}$$
$$= \frac{p'^2 - p^2 + p'(1 - p') - p(1 - p)}{p' - p} = \frac{p' - p}{p' - p} = 1 = U_g - L_g,$$

because in this case both Y_1 and Y_0 follows Bernoulli distribution.

Combining (i)–(iv), we can conclude that condition (2.2) always holds and has no power to detect the violation.

On the other hand, our testable implication can capture such a violation. Consider

$$\mathbb{E}[YD|P = p] = \mathbb{E}[Y_1|P = p]p = \begin{cases} p & \text{if } p < \frac{1}{2}, \\ p^2 & \text{if } p \geq \frac{1}{2}. \end{cases}.$$

It is apparent that $\mathbb{E}[YD|P = p]$ is not a monotone function of p, and therefore violates our testable implication.

C A Finite sample test

This appendix section considers the case with a finite number of J judges, $j = 1, 2, \cdots, J$, and judge j handles a finite number of n_j defendants. For notation simplicity, we assume $n_j = n_{j'} = n^*$, so that the total number of defendants $n = Jn^*$, but our test can be straightforwardly extended to allow for heterogeneous n_j. For defendant i, let $Z_i \in \{1, 2, \cdots, J\}$ be the identity of the judge who handles his/her case. Let p_j be the propensity score or stringency measure of judge j, defined as

$$p_j = P(D_i = 1|Z_i = j)$$

We assume a judge treats all his/her defendants independently. In this section, we consider the case where Y is a binary variable, as in Frandsen, Lefgren, and Leslie (2023). Let $W_i^1 = Y_i D_i$ and $W_i^0 = -Y_i(1 - D_i)$, and define $q_j^1 = \mathbb{E}[W_i^1|Z_i = j] = \mathbb{P}(W_i^1 = 1|Z_i = j)$ and $q_j^0 = \mathbb{E}[W_i^0|Z_i = j] = -\mathbb{P}(W_i^0 = -1|Z_i = j)$. Note that because $D_i \geq W_i^1 \geq 0$ and $0 \geq W_i^0 \geq (D_i - 1)$, we have $0 \leq q_j^1 \leq p_j$ and $0 \leq -q_j^0 \leq p_j$.

The judge leniency design would imply that $(p_j - p_{j'})(q_j^d - q_{j'}^d) \geq 0$ for all $j, j' \in \mathcal{J}$ and $d \in \{0, 1\}$, where $\mathcal{J} \equiv \{1, 2, \cdots, J\}$. With this notation, we rewrite the null hypothesis as

$$H_0 : (p_j - p_{j'})(q_j^d - q_{j'}^d) \geq 0 \text{ for all } j, j' \in \mathcal{J} \text{ and } d \in \{0, 1\} \tag{C.1}$$

To implement a test for H_0, we first consider a test $\phi^{j,j'}$ for the null hypothesis of a given pair (j, j') at $\tilde{\alpha}$ level, that is,

$$\mathbb{P}(\phi^{j,j'} = 1|H_0) \leq \tilde{\alpha},$$

Then, we can define the overall test ϕ such that $\phi = 0$ if $\phi^{j,j'} = 0$ for all pairs (j, j'), and $\phi = 1$ otherwise. That is, we reject H_0 if we reject at least one out of $J(J-1)/2$ pairs. If we choose $\tilde{\alpha} = \frac{2\alpha}{J(J-1)}$, then, we can ensure that

$$\mathbb{P}(\phi = 1|H_0) = \mathbb{P}(\cup_{j>j'}\{\phi^{j,j'} = 1\}|H_0) \leq \sum_{j>j'} \mathbb{P}(\phi^{j,j'} = 1|H_0) \leq \frac{J(J-1)}{2}\tilde{\alpha} = \alpha.$$

Now we construct $\phi^{j,j'}$ for the pair (j, j'). Define $\delta_p = p_j - p_{j'}$, $\delta_q^d = q_j^d - q_{j'}^d$. The relevant null hypothesis is $H_0^{j,j'} : \delta_p \delta_q^d \geq 0$ for $d = 0, 1$. The idea of constructing $\phi^{j,j'}$ is as follows. We first construct the least favorable confidence interval for δ_p and δ_q^d, $d = 0, 1$. Then, suppose we observe that the upper bound of the confidence interval for δ_p is below zero, while the lower bound of the confidence interval for δ_q^d is above zero. In that case, we consider this as evidence against the null hypothesis that δ_p and δ_q^d have to have the same sign. Similarly, we also reject when the lower bound of the confidence interval for δ_p is above zero while the upper bound for δ_q^d is below zero.

Let $\tilde{\tilde{\alpha}} = \frac{\tilde{\alpha}}{4}$. Let $\hat{\delta}_p = \hat{p}_j - \hat{p}_{j'}$ and $\hat{\delta}_q^d = \hat{q}_j^d - \hat{q}_{j'}^d$ be estimators for δ_p and δ_q^d, respectively, where

$$\hat{p}_j = \frac{\sum_{i=1}^N D_i 1\{Z_i = j\}}{n^*}, \quad \hat{q}_j^d = \frac{\sum_{i=1}^N W_i^d 1\{Z_i = j\}}{n^*}.$$

Let $\hat{c}_p$ be the smallest support point of $\hat{\delta}_p$ such that $\mathbb{P}(\hat{\delta}_p > \hat{c}_p|p_j = p_{j'} = 0.5) \leq \tilde{\tilde{\alpha}}$. Note that the distribution of $\hat{\delta}_p$ is symmetric around zero under $p_j = p_{j'} = 0.5$ because defendants handled by judges j and j' are independent, then we would know that $-\hat{c}_p$ is the largest support point of $\hat{\delta}_p$ such that $\mathbb{P}(\hat{\delta}_p < -\hat{c}_p|p_j = p_{j'} = 0.5) \leq \tilde{\tilde{\alpha}}$. Clearly, $\hat{c}_p$ is known and can be tabulated for each n and $\tilde{\tilde{\alpha}}$ by simulation (When n_j and $n_{j'}$ are different, we can simulate the $\tilde{\tilde{\alpha}}$ and $1-\tilde{\tilde{\alpha}}$ quantiles too). Another observation is that for all $\tilde{\tilde{\alpha}} < 0.5$,

$$\mathbb{P}(\hat{\delta}_p < -\hat{c}_p|p_j = p_{j'} = p) \leq \mathbb{P}(\hat{\delta}_p < -\hat{c}_p|p_j = p_{j'} = 0.5) = \tilde{\tilde{\alpha}}, \quad \forall p \in (0, 1).$$

That is, the distribution of $\hat{\delta}_p$ is most dispersed when $p_j = p_{j'} = 0.5$. For instance, if $p_j = p_{j'} = 1$, then $\hat{\delta}_p \equiv 0$ is degenerate. Let $\widehat{CL}_{p,U} = \hat{\delta}_p + \hat{c}_p$ and $\widehat{CL}_{p,L} = \hat{\delta}_p - \hat{c}_p$, then define

$$\widehat{CS}_p \equiv [\widehat{CS}_{p,L}, \widehat{CS}_{p,U}]. \tag{C.2}$$

It is easy to verify that $\widehat{CS}_p$ is a valid $1-2\tilde{\tilde{\alpha}}$ level confidence set for δ_p among the models with $p_j = p_{j'}$.

Similarly, we define $\hat{c}_q^d$ be the largest support point of $\hat{\delta}_q^d$ such that $\mathbb{P}(\hat{\delta}_q^d > \hat{c}_q^d | p_j = p_{j'} = 0.5) \leq \tilde{\tilde{\alpha}}$, and define its $1-2\tilde{\tilde{\alpha}}$ level confidence set as

$$\widehat{CS}_q^d \equiv [\widehat{CS}_{q,L}^d, \widehat{CS}_{q,U}^d], \tag{C.3}$$

where $\widehat{CS}_{q,L}^d = \hat{\delta}_q^d - \hat{c}_q^d$ and $\widehat{CS}_{q,U}^d = \hat{\delta}_q^d + \hat{c}_q^d$.

Now we are ready to define $\phi^{j,j'}$. We reject $H_0^{j,j'}$ when any of the following events happen:

$$\{\widehat{CS}_{p,U} < 0, \widehat{CS}_{q,L}^1 > 0\}, \{\widehat{CS}_{p,U} < 0, \widehat{CS}_{q,L}^0 > 0\}, \{\widehat{CS}_{p,L} > 0, \widehat{CS}_{q,U}^1 < 0\}, \{\widehat{CS}_{p,L} > 0, \widehat{CS}_{q,U}^1 < 0\}$$

We first verify that $\mathbb{P}(\phi^{j,j'} = 1 | H_0^{j,j'}) \leq 4\tilde{\tilde{\alpha}} = \tilde{\alpha}$. Note that,

$$\begin{aligned}\mathbb{P}(\phi^{j,j'} = 1 | H_0^{j,j'}) \leq\ & \mathbb{P}(\{\widehat{CS}_{p,U} < 0, \widehat{CS}_{q,L}^1 > 0\} | H_0^{j,j'}) + \mathbb{P}(\{\widehat{CS}_{p,U} < 0, \widehat{CS}_{q,L}^0 > 0\} | H_0^{j,j'}) \\ & + \mathbb{P}(\{\widehat{CS}_{p,L} > 0, \widehat{CS}_{q,U}^1 < 0\} | H_0^{j,j'}) + \mathbb{P}(\{\widehat{CS}_{p,L} > 0, \widehat{CS}_{q,U}^1 < 0\} | H_0^{j,j'}). \end{aligned} \tag{C.4}$$

Consider the first term on the right-hand side of Equation (C.4). We have,

$$\mathbb{P}(\{\widehat{CS}_{p,U} < 0, \widehat{CS}_{q,L}^1 > 0\} | H_0^{j,j'}) \leq \min\{\mathbb{P}(\{\widehat{CS}_{p,U} < 0\} | H_0^{j,j'}), \mathbb{P}(\widehat{CS}_{q,L}^1 > 0\} | H_0^{j,j'})\}. \tag{C.5}$$

If $H_0^{j,j'}$ is such that $\delta_p > 0$ and $\delta_q^1 > 0$, then,

$$\begin{aligned}\min\{\mathbb{P}(\{\widehat{CS}_{p,U} < 0\} | H_0^{j,j'}), \mathbb{P}(\widehat{CS}_{q,L}^1 > 0\} | H_0^{j,j'})\} &\leq \mathbb{P}(\{\widehat{CS}_{p,U} < 0\} | \delta_p > 0, \delta_q^1 > 0) \\ &\leq \mathbb{P}(\{\widehat{CS}_{p,U} < 0\} | \delta_p = 0) = \mathbb{P}(\{\hat{\delta}_p < -\hat{c}_p\} | \delta_p = 0) = \tilde{\tilde{\alpha}}. \end{aligned} \tag{C.6}$$

where the first equality hold trivially, the first inequality and second equality are by the properties of the confidence interval $\widehat{CS}_p$.

Similarly, if $H_0^{j,j'}$ is such that $\delta_p < 0$ and $\delta_q^1 < 0$, then,

$$\begin{aligned}\min\{\mathbb{P}(\{\widehat{CS}_{p,U} < 0\} | \delta_p < 0, \delta_q^1 < 0), \mathbb{P}(\widehat{CS}_{q,L}^1 > 0\} | H_0^{j,j'})\} &\leq \mathbb{P}(\{\widehat{CS}_{q,L}^1 > 0\} | \delta_p < 0, \delta_q^1 < 0) \\ &\leq \mathbb{P}(\{\widehat{CS}_{q,L}^1 > 0\} | \delta_q^1 = 0) = \mathbb{P}(\{\hat{\delta}_q > \hat{c}_q^1\} | \delta_q^1 = 0) = \tilde{\tilde{\alpha}}. \end{aligned} \tag{C.7}$$

Therefore, we can conclude that the first right-hand side term of Equation (C.4) satisfies

$$\mathbb{P}(\{\widehat{CS}_{p,U} < 0, \widehat{CS}_{q,L}^1 > 0\} | H_0^{j,j'}) \leq \tilde{\tilde{\alpha}}.$$

Applying the same derivations to the remaining three right-hand side terms, we can conclude that

$$\mathbb{P}(\phi^{j,j'} = 1 | H_0^{j,j'}) \leq 4\tilde{\tilde{\alpha}} = \tilde{\alpha}.$$

We summarize the procedure below.

Algorithm C.1 *Finte Sample Test.*

1. *Let J be the number of judges and α be the prechosen significance level. Set $\tilde{\alpha} = \frac{2\alpha}{J(J-1)}$ and $\tilde{\tilde{\alpha}} = \frac{\tilde{\alpha}}{4}$.*

2. *For each judge j, calculate $\hat{\delta}_p = \hat{p}_j - \hat{p}_{j'}$, and $\hat{\delta}_q^d = \hat{q}_j^d - \hat{q}_{j'}^d$, where $(\hat{p}_j, \hat{q}_j^d)$ are sample frequency estimators for (p_j, q_j^d).*

3. *Let B be a large integer (can be millions). For each $b = 1, 2, \cdots, B$, draw two independent random samples of Bernoulli(0.5) random variables, each with sample size n^*. Calculate Δ_b as the difference of the average of the two samples for iteration b. Let $\hat{c}$ be the smallest point from $\{-1, -\frac{n^*-1}{n^*}, \cdots, -\frac{1}{n^*}, 0, \frac{1}{n^*}, \cdots, \frac{n^*-1}{n^*}, 1\}$ such that $\frac{1}{B}\sum_{b=1}^{B} 1\{\Delta_b > \hat{c}\}) \leq \tilde{\tilde{\alpha}}$.*

4. *Set $\hat{c}_p = \hat{c}_q^1 = \hat{c}_q^0 = \hat{c}$.*

5. *Calculate the confidence sets according to Equations (C.2) and (C.3).*

6. *For a given pair (j, j'), set $\phi^{j,j'} = 1$ if any of the following events happen: $\{\widehat{CS}_{p,U} < 0, \widehat{CS}_{q,L}^1 > 0\}, \{\widehat{CS}_{p,U} < 0, \widehat{CS}_{q,L}^0 > 0\}, \{\widehat{CS}_{p,L} > 0, \widehat{CS}_{q,U}^1 < 0\}, \{\widehat{CS}_{p,L} > 0, \widehat{CS}_{q,U}^1 < 0\}$.*

7. *Reject the null hypothesis if $\phi^{j,j'} = 1$ for at least one pair (j, j').*

We report the rejection probability of the finite sample for the design in Section 4.1.2, where we set $\delta_3 = -0.5$. We can see that the power is lower than the asymptotic test that we reported in Table 1, particularly when the violation is relatively mild. This is not surprising because the asymptotic test is based on the assumption that the propensity score is consistently estimated, and therefore, we can consistently estimate the ranking of the propensity score. In contrast, for the finite sample test, the ranking of the propensity score is unknown. Another loss of power is that we only consider one type of interval that $1\{Y \geq 0.5\}$ here, whereas the asymptotic test considers all possible intervals of the form $1\{y \leq Y < y'\}$. However, we still observe that for each given sample size, the rejection frequencies increase as the magnitude of the violation increases, as well as with the number of cases per judge.

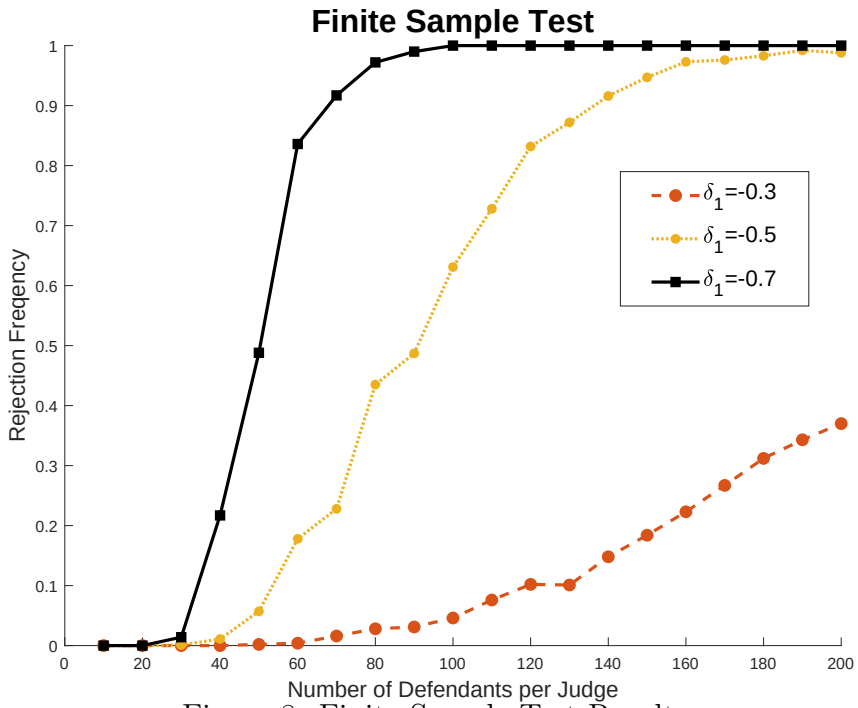


Figure 8: Finite Sample Test Results

To conclude this section, we want to emphasize that while there is a potential loss of power for our test, we trade this off for a substantial computational advantage. As discussed in Frandsen, Lefgren, and Leslie (2023, Supplementary material, page 9), implementing a finite sample test can be quite computationally challenging when involving large-dimensional nonlinear optimization.[15] On the contrary, our test requires little more than drawing Bernoulli random numbers and is very easy to implement. It thus serves as a useful complement to the existing literature.

D Lemmas and Intermediary Results

D.1 Lemmas for the proof of Theorem 2

This section collects useful Lemmas, intermediary results, and additional assumptions for establishing the asymptotic results in Theorem 2.

[15]For this reason, we do not offer a simulation comparison with FLL's finite sample test, for which FLL does not provide a complete simulation study either.

Lemma D.1 *Suppose Assumptions B.2 and B.3 are satisfied, then uniformly in $\ell \in \mathcal{L}$,*

$$\begin{aligned}
&\sqrt{n}(\hat{m}_1(y, r_y, p, r_p, \hat{\theta}) - m_1(y, r_y, p, r_p, \theta_0)) \\
=&\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi_{m_1,i}(y, r_y, p, r_p, \theta_0) + o_p(1) \\
\equiv&\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (m_{1,i}(y, r_y, p, r_p, \theta_0) - m_1(y, r_y, p, r_p, \theta_0) + \nabla_\theta m_1(y, r_y, p, r_p, \theta_0) \cdot s(D_i, Z_i, \theta_0)) + o_p(1).
\end{aligned} \tag{D.1}$$

$$\begin{aligned}
&\sqrt{n}(\hat{m}_0(y, r_y, p, r_p, \hat{\theta}) - m_0(y, r_y, p, r_p, \theta_0)) \\
=&\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi_{m_0,i}(y, r_y, p, r_p, \theta_0) + o_p(1) \\
\equiv&\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (m_{0,i}(y, r_y, p, r_p, \theta_0) - m_0(y, r_y, p, r_p, \theta_0) + \nabla_\theta m_0(y, r_y, p, r_p, \theta_0) \cdot s(D_i, Z_i, \theta_0)) + o_p(1),
\end{aligned} \tag{D.2}$$

$$\begin{aligned}
&\sqrt{n}(\hat{w}(p, r_p, \hat{\theta}) - w(p, r_p, \theta_0)) \\
=&\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi_{w,i}(p, r_p, \theta_0) + o_p(1) \\
\equiv&\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (w_i(p, r_p, \theta_0) - w(p, r_p, \theta_0) + \nabla_\theta w(p, r_p, \theta_0) \cdot s(D_i, Z_i, \theta_0)) + o_p(1)
\end{aligned} \tag{D.3}$$

where functions m_d and w are defined in Equations (A.2) to (A.4) and

$$\begin{aligned}
m_{1i}(y, r_y, p, r_p, \theta) &= D_i 1(y \le Y_i \le y + r_y) 1(p \le P(Z_i, \theta) \le p + r_p), \\
m_{0i}(y, r_y, p, r_p, \theta) &= (D_i - 1) 1(y \le Y_i \le y + r_y) 1(p \le P(Z_i, \theta) \le p + r_p), \\
w_i(p, r_p, \theta) &= 1(p \le P(Z_i, \theta) \le p + r_p).
\end{aligned}$$

Proof. *Let $f_P(p)$ denote the density function of $P(Z; \theta_0)$. Following Hsu and Lieli (2021), we calculate the derivatives for $m_d(y, r_y, p, r_p, \cdot)$ and $w(p, r_p, \cdot)$ as:*

$$\begin{aligned}
\nabla_\theta m_1(y, r_y, p, r_p, \theta_0) = &\mathbb{E}[D 1(y \le Y \le y + r_y) | P(Z, \theta_0) = p] \cdot f_P(p) \mathbb{E}[\nabla_\theta P(Z, \theta_0) | P(Z, \theta_0) = p] \\
&- \mathbb{E}[D 1(y \le Y \le y + r_y) | P(Z, \theta_0) = p + r_p] \cdot f_P(p + r_p) \mathbb{E}[\nabla_\theta P(Z, \theta_0) | P(Z, \theta_0) = p + r_p],
\end{aligned}$$

$$\nabla_\theta m_0(y, r_y, p, r_p, \theta_0) = \mathbb{E}[(D-1)1(y \le Y \le y+r_y)|P(Z,\theta_0) = p]\cdot f_P(p)\mathbb{E}[\nabla_\theta P(Z,\theta_0)|P(Z,\theta_0) = p]$$
$$- \mathbb{E}[(D-1)1(y \le Y \le y+r_y)|P(Z,\theta_0) = p+r_p]\cdot f_P(p+r_p)\mathbb{E}[\nabla_\theta P(Z,\theta_0)|P(Z,\theta_0) = p+r_p],$$

$$\nabla_\theta w(p, r_p, \theta_0) = f_P(p)\mathbb{E}[\nabla_\theta P(Z,\theta_0)|P(Z,\theta_0) = p] - f_P(p+r_p)\mathbb{E}[\nabla_\theta P(Z,\theta_0)|P(Z,\theta_0) = p+r_p].$$

Now we prove Equation (D.1), the results for Equations (D.2) and (D.3) are similar. Note that

$$\begin{aligned}
&\sqrt{n}(\hat{m}_1(y, r_y, p, r_p, \hat{\theta}) - m_1(y, r_y, p, r_p, \theta_0)) \\
=&\sqrt{n}(\hat{m}_1(y, r_y, p, r_p, \hat{\theta}) - m_1(y, r_y, p, r_p, \hat{\theta})) + \sqrt{n}(m_1(y, r_y, p, r_p, \hat{\theta}) - m_1(y, r_y, p, r_p, \theta_0)) \\
=&\sqrt{n}(\hat{m}_1(y, r_y, p, r_p, \hat{\theta}) - m_1(y, r_y, p, r_p, \hat{\theta})) + \nabla_\theta m_1(y, r_y, p, r_p, \theta_0)'\sqrt{n}(\hat{\theta} - \theta_0) + o(\sqrt{n}\|\hat{\theta} - \theta_0\|) \\
=&\sqrt{n}(\hat{m}_1(y, r_y, p, r_p, \hat{\theta}) - m_1(y, r_y, p, r_p, \hat{\theta})) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \nabla_\theta m_1(y, r_y, p, r_p, \theta_0)s(D_i, Z_i, \theta_0) + o_p(1)
\end{aligned} \tag{D.4}$$

where the second equality holds because $m_1(\ell, \theta)$ is continuously differentiable in θ under Assumption B.2-(2), and the third equality is due to Assumption B.3.

Let $\hat{\mathbb{G}}_{m_1}(\theta, \ell) \equiv \sqrt{n}(\hat{m}_1(y, r_y, p, r_p, \theta) - m_1(y, r_y, p, r_p, \theta))$, $\theta \in \Theta, \ell \in \mathcal{L}$. It remains to show that $\sup_{\ell\in\mathcal{L}} |\hat{\mathbb{G}}_{m_1}(\hat{\theta}, \ell) - \hat{\mathbb{G}}_{m_1}(\theta_0, \ell)| = o_p(1)$.

By Assumption B.2-(ii), the class of functions $\{1(p \le P(Z,\theta) \le p + r_p) : \theta \in \Theta, p \in [0,1], r_p \in [0,1]\}$ is a Vapnik-Chervonenkis (VC) class of function. Therefore, the class of functions $\{1\{y \le Y \le y + r_y\} \times 1(p \le P(Z,\theta) : \theta \in \Theta, p \in [0,1], r_p \in [0,1], r_y \in [0,1]\}$ is also VC class. Hence, the process $\hat{\mathbb{G}}_{m_1}$ is stochastically equicontinuous with respect to (θ, ℓ). Note $\hat{\theta} \xrightarrow{p} \theta_0$, then there exist $\delta_n \downarrow 0$ such that with probability approaching one, $(\hat{\theta}, \ell) \in B((\theta_0, \ell), \delta_n)$, where $B((\theta_0, \ell), \delta_n)$ is a ball in $\Theta \times \mathcal{L}$ centered at (θ_0, ℓ) with radius δ_n. Therefore,

$$\begin{aligned}
&\sup_{\ell\in\mathcal{L}} |\sqrt{n}(\hat{m}_1(y, r_y, p, r_p, \hat{\theta}) - m_1(y, r_y, p, r_p, \hat{\theta})) - \sqrt{n}(\hat{m}_1(y, r_y, p, r_p, \theta_0) - m_1(y, r_y, p, r_p, \theta_0))| \\
=&\sup_{\ell\in\mathcal{L}} |\hat{\mathbb{G}}_{m_1}(\hat{\theta}, \ell) - \hat{\mathbb{G}}_{m_1}(\theta_0, \ell)| \\
\le& \sup_{\theta_0\in\Theta, \ell\in\mathcal{L}} \sup_{(\theta',\ell')\in B((\theta_0,\ell),\delta_n)} |\hat{\mathbb{G}}_{m_1}(\theta', \ell') - \hat{\mathbb{G}}_{m_1}(\theta_0, \ell)| = o_p(1).
\end{aligned} \tag{D.5}$$

where the last equality is by the stochastic equicontinuity of the process $\hat{\mathbb{G}}_{m_1}$. Combine both Equations (D.4) and (D.5), the result then follows. □

Lemma D.2 *Suppose Assumptions 2.1 to 2.4, B.2 and B.3 are satisfied, then uniform in ℓ,*

$$\sqrt{n}(\hat{\nu}_1(y, r_y, p_1, p_2, r_p, \hat{\theta}) - \nu_1(y, r_y, p_1, p_2, r_p, \theta_0)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi_{\nu_1,i}(y, r_y, p_1, p_2, r_p, \theta_0) + o_p(1), \tag{D.6}$$

$$\sqrt{n}(\hat{\nu}_0(y, r_y, p_1, p_2, r_p, \hat{\theta}) - \nu_0(y, r_y, p_1, p_2, r_p, \theta_0)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi_{\nu_0,i}(y, r_y, p_1, p_2, r_p, \theta_0) + o_p(1), \tag{D.7}$$

where

$$\begin{aligned}
&\phi_{\nu_1,i}(y, r_y, p_1, p_2, r_p, \theta_0) = w(p_1, r_p, \theta_0) \cdot \phi_{m_1,i}(y, r_y, p_2, r_p, \theta_0) + m_1(y, r_y, p_2, r_p, \theta_0) \cdot \phi_{w,i}(p_1, r_p, \theta_0) \\
&- w(p_2, r_p, \theta_0) \cdot \phi_{m_1,i}(y, r_y, p_1, r_p, \theta_0) - m_1(y, r_y, p_1, r_p, \theta_0) \cdot \phi_{w,i}(p_2, r_p, \theta_0), \\
&\phi_{\nu_0,i}(y, r_y, p_1, p_2, , r_p, \theta_0) = w(p_1, r_p, \theta_0) \cdot \phi_{m_0,i}(y, r_y, p_2, r_p, \theta_0) + m_0(y, r_y, p_2, r_p, \theta_0) \cdot \phi_{w,i}(p_1, r_p, \theta_0) \\
&- w(p_2, r_p, \theta_0) \cdot \phi_{m_0,i}(y, r_y, p_1, r_p, \theta_0) - m_0(y, r_y, p_1, r_p, \theta_0) \cdot \phi_{w,i}(p_2, r_p, \theta_0).
\end{aligned}$$

Furthermore,

$$\sqrt{n}(\widehat{\nu}_1(\cdot, \widehat{\theta}) - \nu_1(\cdot, \theta_0)) \Rightarrow \Phi_{\nu_1}(\cdot), \qquad \sqrt{n}(\widehat{\nu}_0(\cdot, \widehat{\theta}) - \nu_0(\cdot, \theta_0)) \Rightarrow \Phi_{\nu_0}(\cdot),$$

where $\Phi_{\nu_1}(\cdot)$ and $\Phi_{\nu_0}(\cdot)$ are Gaussian processes with variance-covariance kernel generated by $\phi_{\nu_1}(\cdot, \theta_0)$ and $\phi_{\nu_0}(\cdot, \theta_0)$, respectively.

Proof. *We show Equation (D.6). Equation (D.7) holds analogously. Recall*

$$\hat{\nu}_1(\ell) = \hat{m}_1(y, r_y, p_2, r_p, \hat{\theta}) \cdot \hat{w}(p_1, r_p, \hat{\theta}) - \hat{m}_1(y, r_y, p_1, r_p, \hat{\theta}) \cdot \hat{w}(p_2, r_p, \hat{\theta})$$

To save space, for generic ℓ, we write $\hat{m}_1(\hat{\theta}) \equiv \hat{m}_1(\ell, \hat{\theta})$ and $\hat{w}(\hat{\theta}) \equiv \hat{w}(\ell, \hat{\theta})$. Similarly, $m_1(\theta_0) \equiv m_1(\ell, \theta_0)$ and $w(\theta_0) \equiv w(\ell, \theta_0)$. Then,

$$\begin{aligned}
\hat{m}_1(\hat{\theta})\hat{w}(\hat{\theta}) - m_1(\theta_0)w(\theta_0) &= (\hat{m}_1(\hat{\theta}) - m_1(\theta_0) + m_1(\theta_0))(\hat{w}(\hat{\theta}) - w(\theta_0) + w(\theta_0)) - m_1(\theta_0)w(\theta_0) \\
&= (\hat{m}_1(\hat{\theta}) - m_1(\theta_0))w(\theta_0) + (\hat{w}(\hat{\theta}) - w(\theta_0))m_1(\theta_0) + (\hat{m}_1(\hat{\theta}) - m_1(\theta_0))(\hat{w}(\hat{\theta}) - w(\theta_0)) \\
&= (\hat{m}_1(\hat{\theta}) - m_1(\theta_0))w(\theta_0) + (\hat{w}(\hat{\theta}) - w(\theta_0))m_1(\theta_0) + o_p\left(\frac{1}{\sqrt{n}}\right),
\end{aligned}$$

where the last equality is because $\hat{m}_1(\hat{\theta}) - m_1(\theta_0) = O_p(1/\sqrt{n})$ and $\hat{w}(\hat{\theta}) - w(\theta_0) = O_p(1/\sqrt{n})$

by Lemma D.1. Then we have

$$\begin{aligned}\hat{\nu}_1(\ell) - \nu_1(\ell) =& w(p_1, r_p, \theta_0) \cdot (\hat{m}_1(y, r_y, p_2, r_p, \hat{\theta}) - m_1(y, r_y, p_2, r_p, \theta_0)) \\ &+ m_1(y, r_y, p_2, r_p, \theta_0) \cdot (\hat{w}(p_1, r_p, \hat{\theta}) - w(p_1, r_p, \theta_0)) \\ &- w(p_2, r_p, \theta_0) \cdot (\hat{m}_1(y, r_y, p_1, r_p, \hat{\theta}) - m_1(y, r_y, p_1, r_p, \theta_0)) \\ &- m_1(y, r_y, p_1, r_p, \theta_0) \cdot (\hat{w}(p_2, r_p, \hat{\theta}) - w(p_2, r_p, \theta_0)) + o_p\left(\frac{1}{\sqrt{n}}\right).\end{aligned}$$

Equation (D.6) then follows by inserting Equations (D.1) to (D.3) to the above equation.

Finally, under Assumption B.2, each element of $\nabla_\theta m_1(y, r_y, p, r_p, \theta_0)$ is Lipschitz continuous in y, r_y, p, r_p and it implies that $\{\partial m_1(y, r_y, p, r_p, \theta_0)/\partial\theta_j : (y, r_y, p, r_p) \in [0,1]^4\}$ is a VC class of functions for each j. Similarly, each element of $\nabla_\theta w(p, r_p, \theta_0)$ is Lipschitz continuous in p, r_p. It follows that $\{\phi_{m_1}(y, r_y, p, r_p, \theta_0) : (y, r_y, p, r_p) \in [0,1]^4\}$, $\{\phi_{m_0}(y, r_y, p, r_p, \theta_0) : (y, r_y, p, r_p) \in [0,1]^4\}$ and $\{\phi_w(p, r_p, \theta_0) : (p, r_p) \in [0,1]^2\}$ are all VC classes of functions. weak convergence follows from the fact that $\{\phi_{\nu_0}(y, r_y, p_1, p_2, , r_p, \theta_0) : (y, r_y, p_1, p_2, r_p) \in [0,1]^5\}$ and $\{\phi_{\nu_0}(y, r_y, p_1, p_2, , r_p, \theta_0) : (y, r_y, p_1, p_2, r_p) \in [0,1]^5\}$ are both VC classes of functions. Therefore, we have

$$\sqrt{n}(\widehat{\nu}_1(\cdot, \widehat{\theta}) - \nu_1(\cdot, \theta_0)) \Rightarrow \Phi_{\nu_1}(\cdot), \quad \sqrt{n}(\widehat{\nu}_0(\cdot, \widehat{\theta}) - \nu_0(\cdot, \theta_0)) \Rightarrow \Phi_{\nu_0}(\cdot).$$

□

Lemma D.3 *Suppose Assumptions 2.1 to 2.4, B.2, B.3 and B.5 are satisfied, then uniform in ℓ over $\mathcal{L}$,*

$$\begin{aligned}&\sqrt{n}(\hat{\nu}_1^b(y, r_y, p_1, p_2, r_p, \hat{\theta}^b) - \hat{\nu}_1(y, r_y, p_1, p_2, r_p, \hat{\theta})) \\ =&\frac{1}{\sqrt{n}}\sum_{i=1}^n (W_i - 1)\phi_{\nu_1,i}(y, r_y, p_1, p_2, r_p, \theta_0) + o_p(1), \qquad \text{(D.8)} \\ &\sqrt{n}(\hat{\nu}_0^b(y, r_y, p_1, p_2, r_p, \hat{\theta}^b) - \hat{\nu}_0(y, r_y, p_1, p_2, r_p, \hat{\theta})) \\ =&\frac{1}{\sqrt{n}}\sum_{i=1}^n (W_i - 1)\phi_{\nu_0,i}(y, r_y, p_1, p_2, r_p, \theta_0) + o_p(1), \qquad \text{(D.9)}\end{aligned}$$

where $\phi_{\nu_1,i}(y, r_y, p_1, p_2, r_p, \theta_0)$ and $\phi_{\nu_0,i}(y, r_y, p_1, p_2, , r_p, \theta_0)$ are the same as in Lemma D.2.

The proof to Lemma D.3 is similar to Lemma D.2 and is therefore omitted.

Lemma D.4 *Suppose Assumptions 2.1 to 2.4, B.2, B.3 and B.5 are satisfied, then $\hat{\sigma}_d^2(\ell)$ defined in (3.6) satisfies that for $d = 0, 1$, $\sup_\ell |\hat{\sigma}_d^2(\ell) - \sigma_d^2(\ell)| = o_p(1)$.*

Proof. *Recall that for a given $\ell \in \mathcal{L}$,*

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

It can be written as

$$\hat{\sigma}_d^2(\ell) = \frac{n}{B}\sum_{b=1}^{B}\left(\hat{\nu}_d^b(\ell) - \hat{\nu}_d(\ell)\right)^2 + 2\frac{n}{B}\sum_{b=1}^{B}\left(\hat{\nu}_d^b(\ell) - \hat{\nu}_d(\ell)\right)\left(\hat{\nu}_d(\ell) - \overline{\hat{\nu}_d^b}(\ell)\right) + \frac{n}{B}\sum_{b=1}^{B}\left(\hat{\nu}_d(\ell) - \overline{\hat{\nu}_d^b}(\ell)\right)^2 \tag{D.10}$$

We first consider the second term on the right-hand side of Equation (D.10). Let $\bar{W}_i = \frac{1}{B}\sum_{b=1}^{B} W_i^b$, Using Lemma D.3, we know that for a given $b = 1, 2, \cdots, B$, and uniformly over $\ell \in \mathcal{L}$,

$$\hat{\nu}_d^b(\ell) - \hat{\nu}_d(\ell) = \frac{1}{n}\sum_{i=1}^{n}(W_i^b - 1)\phi_{\nu_d,i}(\ell, \theta_0) + o_p(1).$$

So it can be written as

$$\begin{aligned}
&\frac{n}{B}\sum_{b=1}^{B}\left(\hat{\nu}_d^b(\ell) - \hat{\nu}_d(\ell)\right)\left(\hat{\nu}_d(\ell) - \overline{\hat{\nu}_d^b}(\ell)\right) \\
=&\frac{1}{B}\frac{1}{n}\sum_{b=1}^{B}\Big(\sum_{i=1}^{n}(W_i^b - 1)\phi_{\nu_d,i}(\ell, \theta_0)\Big)\Big(\sum_{i=1}^{n}(\bar{W}_i - 1)\phi_{\nu_d,i}(\ell, \theta_0)\Big) + o_p(1) \\
=&\frac{1}{B}\frac{1}{n}\sum_{b=1}^{B}\sum_{i=1}^{n}(W_i^b - 1)(\bar{W}_i - 1)\phi_{\nu_d,i}^2(\ell, \theta_0) + \frac{1}{B}\frac{1}{n}\sum_{b=1}^{B}\sum_{i\neq j}^{n}(W_i^b - 1)(\bar{W}_j - 1)\phi_{\nu_d,i}(\ell, \theta_0)\phi_{\nu_d,j}(\ell, \theta_0) + o_p(1) \\
=&\frac{1}{B^2}\frac{1}{n}\sum_{b=1}^{B}\sum_{i=1}^{n}(W_i^b - 1)^2\phi_{\nu_d,i}^2(\ell, \theta_0) + \frac{1}{B^2}\frac{1}{n}\sum_{b=1}^{B}\sum_{b'\neq b}^{B}\sum_{i=1}^{n}(W_i^b - 1)(W_j^{b'} - 1)\phi_{\nu_d,i}^2(\ell, \theta_0) \\
&+ \frac{1}{B}\frac{1}{n}\sum_{b=1}^{B}\sum_{i\neq j}^{n}(W_i^b - 1)(\bar{W}_j - 1)\phi_{\nu_d,i}(\ell, \theta_0)\phi_{\nu_d,j}(\ell, \theta_0) + o_p(1)
\end{aligned}$$

The first right-hand side term is of order $\frac{1}{B}$ and is negligible as $B \to \infty$. The second term on the right-hand side is negligible because $E[(W_i^b - 1)(W_i^{b'} - 1)|(Y, D, Z)] = 0$ as long as $b \neq b'$. The third term on the right-hand side is negligible because $E[(W_i^b - 1)(W_j^b - 1)|(Y, D, Z)] = 0$ as long as $i \neq j$. For similarly reasoning, the third right-hand side term of Equation (D.10) is also negligible as $B \to \infty$.

Now consider the first term on the right-hand side of Equation (D.10). Uniformly over ℓ,

$$\frac{n}{B}\sum_{b=1}^{B}\left(\hat{\nu}_d^b(\ell)-\hat{\nu}_d(\ell)\right)^2=\frac{1}{B}\frac{1}{n}\sum_{b=1}^{B}\left(\sum_{i=1}^{n}(W_i^b-1)\phi_{\nu_d,i}(\ell,\theta_0)\right)^2+o_p(1)$$
$$=\frac{1}{B}\frac{1}{n}\sum_{b=1}^{B}\sum_{i=1}^{n}(W_i^b-1)^2\phi_{\nu_d,i}^2(\ell,\theta_0)+\frac{1}{B}\frac{1}{n}\sum_{b=1}^{B}\sum_{i=1}^{n}\sum_{j\neq i}(W_i^b-1)(W_j^b-1)\phi_{\nu_d,i}(\ell,\theta_0)\phi_{\nu_d,j}(\ell,\theta_0)+o_p(1).$$

Conditioning on the sample, because W_i^b are i.i.d. across b and i, has expectation and variance equal to one, we know $E[(W_i^b-1)(W_j^b-1)|(Y,D,Z)]=0$ and $E[(W_i^b-1)^2|(Y,D,Z)]=1$. As $B\to\infty$, the right-hand side converges in probability (with respect to the distribution of $\{W^b\}_{b=1}^B$) to $\frac{1}{n}\sum_{i=1}^n\phi_{\nu_d,i}^2(\ell,\theta_0)+o_p(1)$, which in turn converges to $\sigma_d^2(\ell)$ uniformly over ℓ as $n\to\infty$. □

D.2 The influence function with covariate case

In this subsection, we derive the influence function for estimating $\nu_d(\ell)$ in the presence of covariates. First, we estimate $\theta_0\equiv(\theta_{0z},\theta_{0x})$ by MLE,

$$\begin{aligned}\hat{\theta}&=\underset{\theta\in\Theta}{\text{argmax}}\,\frac{1}{n}\sum_{i=1}^{n}\log f(Y_i,D_i,Z_i,X_i,\theta)\\&\equiv\underset{\theta\in\Theta}{\text{argmax}}\,\frac{1}{n}\sum_{i=1}^{n}D_i\log P(Z_i,X_i,\theta)+(1-D_i)\log(1-P(Z_i,X_i,\theta)).\end{aligned}\tag{D.11}$$

where $P(z,x,\theta)$ is parameterized and depends on (z,x) and $\theta\equiv(\theta_z',\theta_x')'$ through $z'\theta_z+x'\theta_x$. For example, $P(z,x,\theta)=\Phi(z'\theta_z+x'\theta_x)$ for Probit or $P(z,x,\theta)=\frac{exp(z'\theta_z+x'\theta_x)}{1+\exp(z'\theta_z+x'\theta_x)}$ for Logit.

As in Appendix D.1, we make the following assumptions.

Assumption D.1 *Assuming following conditions hold*

1. *The conditional density of (Y,X,D) given $P(Z,X,\theta_0)=p$, denoted by $f_{Y,X,D|P}(y,x,d|p)$, is Lipschitz continuous in (y,x,p) over the joint support of (Y,X,P) for $d=0,1$.*

2. *For all $z\in\mathcal{Z}$ and $x\in\mathcal{X}$, $P(z,x,\theta)$ is continuously differentiable in θ at θ_0 with bounded derivatives.*

Assumption D.2 *The estimator $\hat{\theta}$, $\hat{\beta}_1$, $\hat{\beta}_0$ admits an influence function of the following form,*

$$\sqrt{n}(\hat{\theta}-\theta_0)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}s_{\theta_0}(D_i,Z_i,X_i,\theta_0)+o_p(1), \tag{D.12}$$

$$\sqrt{n}(\hat{\beta}_1-\beta_1)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}s_{\beta_1}(D_i,Y_i,Z_i,X_i,\beta_1)+o_p(1), \tag{D.13}$$

$$\sqrt{n}(\hat{\beta}_0-\beta_0)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}s_{\beta_0}(D_i,Y_i,Z_i,X_i,\beta_0)+o_p(1), \tag{D.14}$$

where $s_{\theta_0}(\cdot)$, $s_{\beta_1}(\cdot)$ and $s_{\beta_0}(\cdot)$ are measurable, satisfying $E[s_{\theta_0}(D_i,Z_i,X_i,\theta_0)]=0$,$E[s_{\beta_1}(D_i,Y_i,Z_i,X_i,\beta_1)]=0$, $E[s_{\beta_0}(D_i,Y_i,Z_i,X_i,\beta_0)]=0$, $E[\sup_\theta\|s_{\theta_0}(D_i,Z_i,\theta)\|^{2+\delta}]<\infty$, $E[\sup_\beta\|s_{\beta_1}(D_i,Y_i,Z_i,X_i,\beta)\|^{2+\delta}]<\infty$, and $E[\sup_\beta\|s_{\beta_0}(D_i,Y_i,Z_i,X_i,\beta)\|^{2+\delta}]<\infty$ for some $\delta>0$.

Note that under similar conditions as in Section 4 of Hsu, Liao and Lin (2022, Econometric Reviews), (D.13) and (D.14) would hold. We define the following quantities for generic (y,r_y,p,r_p,b,θ):

$$m_1(y,r_y,p,r_p,b,\theta)=\mathbb{E}[D1(y\le Y-X'b\le y+r_y)1(p\le P(Z,X,\theta)\le p+r_p)], \tag{D.15}$$

$$m_0(y,r_y,p,r_p,b,\theta)=\mathbb{E}[(D-1)1(y\le Y-X'b\le y+r_y)1(p\le P(Z,X,\theta)\le p+r_p)], \tag{D.16}$$

$$w(p,r_p,\theta)=\mathbb{E}[1(p\le P(Z,X,\theta)\le p+r_p)]. \tag{D.17}$$

Let $f_P(p)$ denote the density function of $P(Z,X,\theta_0)\equiv\mathbb{P}(D=1|X,Z;\theta_0)$. Following the calculation in Hsu and Lieli (2021), we can analogously obtain the derivatives with respect to θ, evaluating at the true parameter values $(\beta_1,\beta_0,\theta_0)$ as

$$\begin{aligned}
&\nabla_\theta m_1(y,r_y,p,r_p,\beta_1,\theta_0)\\
=&\mathbb{E}[D1(y\le Y-X'\beta_1\le y+r_y)|P(Z,X,\theta_0)=p]\cdot f_P(p)\mathbb{E}[\nabla_\theta P(Z,X,\theta_0)|P(Z,X,\theta_0)=p]\\
&-\mathbb{E}[D1(y\le Y-X'\beta_1\le y+r_y)|P(Z,X,\theta_0)=p+r_p]\cdot f_P(p+r_p)\mathbb{E}[\nabla_\theta P(Z,X,\theta_0)|P(Z,X,\theta_0)=p+r_p],\\
&\nabla_\theta m_0(y,r_y,p,r_p,\beta_0,\theta_0)\\
=&\mathbb{E}[(D-1)1(y\le Y-X'\beta_0\le y+r_y)|P(Z,X,\theta_0)=p]\cdot f_P(p)\mathbb{E}[\nabla_\theta P(Z,X,\theta_0)|P(Z,X,\theta_0)=p]\\
&-\mathbb{E}[(D-1)1(y\le Y-X'\beta_0\le y+r_y)|P(Z,X,\theta_0)=p+r_p]\cdot f_P(p+r_p)\mathbb{E}[\nabla_\theta P(Z,X,\theta_0)|P(Z,X,\theta_0)=p+r_p],\\
&\nabla_\theta w(p,r_p,\theta_0)\\
=&f_P(p)\mathbb{E}[\nabla_\theta P(Z,X,\theta_0)|P(Z,X,\theta_0)=p]-f_P(p+r_p)\mathbb{E}[\nabla_\theta P(Z,X,\theta_0)|P(Z,X,\theta_0)=p+r_p].
\end{aligned}$$

In addition, let $f_{u_d|zxd}(y|z,x,d)$ denote the conditional density of U_d conditional on $(Z,X,D)=$

(z, x, d), then the derivatives with respect to β, evaluating at the true parameter values $(\beta_1, \beta_0, \theta_0)$ are

$$
\begin{aligned}
&\nabla_\beta m_1(y, r_y, p, r_p, \beta_1, \theta_0) \\
=&\mathbb{E}[P(Z, X, \theta_0)(f_{u_1|zxd}(y + r_y|Z, X, 1) - f_{u_1|zxd}(y|Z, X, 1)X \cdot 1(p \le P(Z, X, \theta) \le p + r_p)]], \\
&\nabla_\beta m_0(y, r_y, p, r_p, \beta_0, \theta_0) \\
=&\mathbb{E}[(1 - P(Z, X, \theta_0))(f_{u_0|zxd}(y + r_y|Z, X, 0) - f_{u_0|zxd}(y|Z, X, 0)X \cdot 1(p \le P(Z, X, \theta) \le p + r_p)]].
\end{aligned}
$$

Let the estimators for $m_1(y, r_y, p, r_p, \beta, \theta)$, $m_0(y, r_y, p, r_p, \beta, \theta)$ and $w(p, r_p, \theta)$ be

$$
\begin{aligned}
\hat{m}_1(y, r_y, p, r_p, \beta, \theta) &= \frac{1}{n}\sum_{i=1}^{n} m_{1,i}(y, r_y, p, r_p, \beta, \theta), \\
\hat{m}_0(y, r_y, p, r_p, \beta, \theta) &= \frac{1}{n}\sum_{i=1}^{n} m_{0,i}(y, r_y, p, r_p, \beta, \theta), \\
\hat{w}(p, r_p, \theta) &= \frac{1}{n}\sum_{i=1}^{n} w_i(p, r_p, \theta).
\end{aligned}
$$

where

$$
\begin{aligned}
&m_{1,i}(y, r_y, p, r_p, \beta, \theta) = D_i 1(y \le Y_i - X_i\beta \le y + r_y)1(p \le P(Z_i, X_i, \theta) \le p + r_p), \\
&m_{0,i}(y, r_y, p, r_p, \beta, \theta) = (1 - D_i)1(y \le Y_i - X_i\beta \le y + r_y)1(p \le P(Z_i, X_i, \theta) \le p + r_p), \\
&w_i(p, r_p, \theta) = 1(p \le P(Z_i, X_i, \theta) \le p + r_p),
\end{aligned}
$$

and

$$
\begin{aligned}
&\sqrt{n}(\hat{m}_1(y, r_y, p, r_p, \hat{\beta}_1, \hat{\theta}) - m_1(y, r_y, p, r_p, \beta_1, \theta_0)) \\
=&\frac{1}{\sqrt{n}}\sum_{i=1}^{n} m_{1,i}(y, r_y, p, r_p, \beta_1, \theta_0) - m_1(y, r_y, p, r_p, \beta_1, \theta_0) + \nabla_\theta m_1(y, r_y, p, r_p, \beta_1, \theta_0) \cdot s(D_i, Z_i, X_i, \theta_0) \\
&\quad + \nabla_\beta m_1(y, r_y, p, r_p, \beta_1, \theta_0) \cdot s_{\beta_1}(D_i, Y_i, Z_i, X_i, \beta_1) + o_p(1) \\
\equiv&\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \phi_{m_1,i}(y, r_y, p, r_p, \beta_1, \theta_0) + o_p(1),
\end{aligned}
$$

$$
\begin{aligned}
&\sqrt{n}(\hat{m}_0(y, r_y, p, r_p, \hat{\beta}_0, \hat{\theta}) - m_0(y, r_y, p, r_p, \beta_0, \theta_0)) \\
=&\frac{1}{\sqrt{n}} \sum_{i=1}^{n} m_{0,i}(y, r_y, p, r_p, \beta_0, \theta_0) - m_0(y, r_y, p, r_p, \beta_0, \theta_0) + \nabla_\theta m_0(y, r_y, p, r_p, \beta_0, \theta_0) \cdot s(D_i, Z_i, X_i, \theta_0) \\
&+ \nabla_\beta m_0(y, r_y, p, r_p, \beta_0, \theta_0) \cdot s_{\beta_0}(D_i, Y_i, Z_i, X_i, \beta_0) + o_p(1) \\
\equiv&\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi_{m_0,i}(y, r_y, p, r_p, \theta_0) + o_p(1),
\end{aligned}
$$

$$
\begin{aligned}
&\sqrt{n}(\hat{w}(p, r_p, \hat{\theta}) - w(p, r_p, \theta_0)) \\
=&\frac{1}{\sqrt{n}} \sum_{i=1}^{n} w_i(p, r_p, \theta_0) - w(p, r_p, \theta_0) + \nabla_\theta w(p, r_p, \theta_0) \cdot s(D_i, Z_i, X_i, \theta_0) + o_p(1) \\
\equiv&\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi_{w,i}(p, r_p, \theta_0) + o_p(1).
\end{aligned}
$$

By Assumption D.1, all elements of $\nabla_\theta m_1(y, r_y, p, r_p, \beta_1, \theta_0)$, $\nabla_\beta m_1(y, r_y, p, r_p, \beta_1, \theta_0)$, $\nabla_\theta m_0(y, r_y, p, r_p, \beta_0, \theta_0)$, and $\nabla_\beta m_0(y, r_y, p, r_p, \beta_0, \theta_0)$, are Lipschitz continuous in y, r_y, p, r_p, and each element of $\nabla_\theta w(p, r_p, \theta_0)$ is Lipschitz continuous in p, r_p. It follows that $\{\phi_{m_1}(y, r_y, p, r_p, \beta_1\theta_0) : (y, r_y, p, r_p) \in [0,1]^4\}$, $\{\phi_{m_0}(y, r_y, p, r_p, \beta_0, \theta_0) : (y, r_y, p, r_p) \in [0,1]^4\}$ and $\{\phi_w(p, r_p, \theta_0) : (p, r_p) \in [0,1]^2\}$ are all VC classes of functions. Finally, let

$$
\begin{aligned}
&\nu_1(y, r_y, p_1, p_2, r_p, \beta_1, \theta_0) = m_1(y, r_y, p_2, r_p, \beta_1, \theta_0) \cdot w(p_1, r_p, \theta_0) - m_1(y, r_y, p_1, r_p, \beta_1, \theta_0) \cdot w(p_2, r_p, \theta_0), \\
&\nu_0(y, r_y, p_1, p_2, r_p, \beta_1, \beta_0, \theta_0) = m_0(y, r_y, p_2, r_p, \beta_0, \theta_0) \cdot w(p_1, r_p, \theta_0) - m_0(y, r_y, p_1, r_p, \beta_0, \theta_0) \cdot w(p_2, r_p, \theta_0), \\
&\hat{\nu}_1(y, r_y, p_1, p_2, r_p, \hat{\beta}_1, \hat{\theta}) = \hat{m}_1(y, r_y, p_2, r_p, \hat{\beta}_1, \hat{\theta}) \cdot \hat{w}(p_1, r_p, \hat{\theta}) - \hat{m}_1(y, r_y, p_1, r_p, \hat{\beta}_1, \hat{\theta}) \cdot \hat{w}(p_2, r_p, \hat{\theta}), \\
&\hat{\nu}_0(y, r_y, p_1, p_2, r_p, \hat{\beta}_0, \hat{\theta}) = \hat{m}_0(y, r_y, p_2, r_p, \hat{\beta}_0, \hat{\theta}) \cdot \hat{w}(p_1, r_p, \hat{\theta}) - \hat{m}_0(y, r_y, p_1, r_p, \hat{\beta}_0, \hat{\theta}) \cdot \hat{w}(p_2, r_p, \hat{\theta}).
\end{aligned}
$$

Lemma D.5 *Suppose Assumptions 2.1 to 2.4, 3.3, D.1 and D.2 are satisfied, then,*

$$
\sqrt{n}(\hat{\nu}_1(y, r_y, p_1, p_2, r_p, \hat{\beta}_1, \hat{\theta}) - \nu_1(y, r_y, p_1, p_2, r_p, \beta_1, \theta_0)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi_{\nu_1,i}(y, r_y, p_1, p_2, r_p, \beta_1, \theta_0) + o_p(1), \tag{D.18}
$$

$$
\sqrt{n}(\hat{\nu}_0(y, r_y, p_1, p_2, r_p, \hat{\beta}_0, \hat{\theta}) - \nu_0(y, r_y, p_1, p_2, r_p, \beta_0, \theta_0)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi_{\nu_0,i}(y, r_y, p_1, p_2, r_p, \beta_0, \theta_0) + o_p(1), \tag{D.19}
$$

where

$$
\begin{aligned}
&\phi_{\nu_1,i}(y, r_y, p_1, p_2, r_p, \beta_1, \theta_0) \\
=&w(p_1, r_p, \theta_0) \cdot \phi_{m_1,i}(y, r_y, p_2, r_p, \beta_1, \theta_0) + m_1(y, r_y, p_2, r_p, \beta_1, \theta_0) \cdot \phi_{w,i}(p_1, r_p, \theta_0) \\
&- w(p_2, r_p, \theta_0) \cdot \phi_{m_1,i}(y, r_y, p_1, r_p, \beta_1, \theta_0) + m_1(y, r_y, p_1, r_p, \beta_1, \theta_0) \cdot \phi_{w,i}(p_2, r_p, \theta_0), \\
&\phi_{\nu_0,i}(y, r_y, p_1, p_2, r_p, \beta_0, \theta_0) \\
=&w(p_1, r_p, \theta_0) \cdot \phi_{m_0,i}(y, r_y, p_2, r_p, \beta_0, \theta_0) + m_0(y, r_y, p_2, r_p, \beta_0, \theta_0) \cdot \phi_{w,i}(p_1, r_p, \theta_0) \\
&- w(p_2, r_p, \theta_0) \cdot \phi_{m_0,i}(y, r_y, p_1, r_p, \beta_0, \theta_0) + m_0(y, r_y, p_1, r_p, \beta_0, \theta_0) \cdot \phi_{w,i}(p_2, r_p, \theta_0).
\end{aligned}
$$

The proofs are similar to those in Appendix D.1, so we omit the details.

E Additional Empirical Results

Table 4: FLL Semi-parametric test, B-Spline

	2 Knots			3 Knots			4 Knots			5 Knots		
	p_f	p_s	combined	p_f	p_s	combined	p_f	p_s	combined	p_f	p_s	combined
All	0.13	1.00	0.25	0.06	1.00	0.11	0.02	1.00	0.04	na	1.00	1.00
Aggressive Assault	0.02	1.00	0.04	0.01	0.91	0.02	0.00	0.96	0.00	na	1.00	1.00
Robbery	0.13	0.73	0.25	0.06	0.99	0.11	0.02	0.44	0.04	na	0.45	0.90
Drug Sale	0.18	0.83	0.36	0.09	0.33	0.18	0.03	0.55	0.06	na	0.73	1.00
Drug Possession	0.45	0.82	0.89	0.31	1.00	0.61	0.14	0.99	0.27	na	0.98	1.00

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 (2024): "On the use of outcome tests for detecting bias in decision making," *Review of Economic Studies*, 91(4), 2135–2167.

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.

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.

Jochmans, K. (2023): "Many (Weak) Judges in Judge-Leniency Designs," .

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.

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.

Li, L., D. Kédagni, and I. Mourifié (2024): "Discordant relaxations of misspecified models," *Quantitative Economics*, 15(2), 331–379.

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.

Mourifié, I., and Y. Wan (2025): "Layered policy analysis in program evaluation using the marginal treatment effect," *Journal of Econometrics*, 251, 106060.

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.

Ren, M. (2024): "Extrapolating LATE with Weak IVs," *working paper*.

Sithole, L. (2024): "A Locally Robust Semiparametric Approach to Examiner IV Designs," *arXiv preprint arXiv:2404.19144*.

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.

Yap, L. (2024): "Inference with Many Weak Instruments and Heterogeneity," *arXiv preprint arXiv:2408.11193*.