Falsifying Marginal Treatment Effects

Minghai Mao* Pedro H. C. Sant'Anna† Xiaojun Song ‡

June 16, 2026

Abstract

The marginal treatment effect (MTE) framework is commonly used to recover policy-relevant treatment effects in settings with selection on unobservables. The credibility of MTE-based conclusions depends on three IV-identifying assumptions (exclusion, random assignment, and monotonicity), yet formal tools for assessing these assumptions in the MTE setting remain underdeveloped. We characterize the sharp, observable implications of these IV-identifying assumptions in the nonparametric MTE model and derive residual-based counterparts to the semiparametric specifications commonly used in applications. Under the maintained structure, these residual restrictions combine index sufficiency with the stochastic monotonicity of outcome-treatment subdistributions in the propensity score. We develop specification tests for these properties while accounting for the propensity score's generated-regressor nature and the outcome residuals. Monte Carlo simulations show good finite-sample performance, and an empirical application finds no statistical evidence against the maintained semiparametric MTE specification in a widely used education dataset.

*Li Anmin Institute of Economic Research, Liaoning University. Financial support from the National Science Foundation of China [Grant Number 72203084] is gratefully acknowledged.

†Department of Economics, Emory University.

‡Guanghua School of Management, Peking University. Financial support from the National Natural Science Foundation of China [Grant Numbers 72373007 and 72333001] is gratefully acknowledged. The author also gratefully acknowledges the research support from the Center for Statistical Science of Peking University, China, and the Key Laboratory of Mathematical Economics and Quantitative Finance (Peking University) of the Ministry of Education, China.

1 Introduction

The marginal treatment effect (MTE) framework has become a central tool for empirical work on treatment effect heterogeneity in settings with selection on unobservables. Applications span economics of education (Carneiro, Heckman and Vytlacil, 2011), economics of crime (Dobbie, Goldin and Yang, 2018), health economics (Gupta, Howell, Yannelis and Gupta, 2024), and many other fields; see Mogstad and Torgovitsky (2024) for a comprehensive survey. A key appeal of the MTE framework is that, beyond documenting heterogeneity, it can be used to inform counterfactual questions about how treatment gains vary across individuals and how marginal policy changes may affect populations that have not yet been treated. It does so by connecting rigorous econometric identification to policy-relevant counterfactuals. The credibility of such exercises, however, depends on the validity of the instrumental variables (IV) used to recover the MTE. When exclusion, random assignment, or monotonicity fail, MTE-based estimates can yield a distorted picture of who benefits from treatment and how policy changes would affect untreated populations. Yet, despite how much applied work relies on these assumptions, the toolkit for formally assessing their plausibility remains underdeveloped, particularly when researchers work with multiple continuous instruments and covariates. This paper narrows this gap by proposing falsification tests to assess the plausibility of the MTE framework.

Our first contribution is to provide a sharp characterization of the testable implications of IV identifying assumptions in MTE models. Building on Heckman and Vytlacil (2005), we show that the full set of observable restrictions consists of two distinct components: *index sufficiency*, requiring that the propensity score captures all the relevant information in the instruments for the outcome distribution, and *stochastic monotonicity* of outcomes and treatment status with respect to the propensity score. Recent work has clarified the sharp content of monotonicity restrictions in settings where the propensity score or judge-leniency ordering is the central object (Carr and Kitagawa, 2021 and Coulibaly, Hsu, Mourifié and Wan, 2024). Our contribution is complementary: in semiparametric MTE applications with multiple instruments and generated propensity scores, index sufficiency is an additional empirically relevant implication, and the two components should be characterized and assessed jointly. We further extend this dual characterization to the semiparametric specifications commonly used in practice, deriving sharp, testable implications for residual distributions under standard linear functional forms. An important practical takeaway is that our framework provides a rigorous way to assess the validity of the exact MTE specifications that applied researchers typically estimate—ensuring that the testing procedure matches the empirical model rather than an idealized benchmark. Importantly, the semiparametric characterization is a sharpness statement for the maintained residual structure: a rejection of the residual restrictions falsifies at least one component of the joint specification (the IV assumptions, the linear residual form, or the propensity score model), while a non-rejection provides support for the specification but does not prove IV validity.

Our second contribution is to develop a unified testing procedure that jointly evaluates index

sufficiency and stochastic monotonicity in MTE models. Simpler approaches to this problem typically rely on coarsening the continuous propensity score into discrete categories and adapting strategies from the LATE literature, such as Kitagawa (2015, K15 hereafter) and Mourifié and Wan (2017). While intuitive, such coarsening discards information and can miss key violations of the original restrictions. Our procedure preserves the full complexity of the continuous setting, but doing so requires overcoming two non-standard statistical challenges.

The first challenge concerns the monotonicity shape-restriction, which is an infinite-dimensional inequality. We adopt a copula transformation that recasts this restriction as a functional equality between the copula and its least concave majorant (LCM), thereby preserving its information content without coarsening the propensity score. This step is technically nontrivial because the LCM operator is not Hadamard differentiable, rendering standard bootstrap procedures invalid. To address this challenge, we rely on the numerical bootstrap of Hong and Li (2018, 2020), building on Fang and Santos (2018). Relative to approaches that are necessarily conservative under broad inequality nulls, our LCM formulation can deliver asymptotically exact size control in the nondegenerate contact-set case while remaining conservative in degenerate cases. The second challenge is that both the propensity score and the outcome residuals are generated regressors. We derive influence function representations that jointly track sampling variation from first-stage parametric estimation and nonparametric estimation of the residual distribution conditional on the propensity score. By addressing these two challenges, our falsification tests provide strong statistical guarantees and can be used to probe the plausibility of MTE-type assumptions in applications.

To illustrate the empirical relevance of our approach, we apply the test to Carneiro et al. (2011), which estimates marginal returns to college using multiple continuous instruments. We find no statistical evidence against index sufficiency, monotonicity, or their joint restriction in this application. To our knowledge, this is the first formal statistical assessment of both index sufficiency and monotonicity as observable implications of the IV identifying assumptions underlying semiparametric MTE analysis. As with any specification test, however, non-rejection provides support for, but cannot confirm, those identifying assumptions.

Related literature: Within the MTE framework, our paper is most closely related to Carr and Kitagawa (2021) and Coulibaly et al. (2024). More broadly, it contributes to the literature that uses sharp observable restrictions to assess IV validity in heterogeneous treatment effect models, including Kitagawa (2015, K15 hereafter), Mourifié and Wan (2017), and Sun (2023) in discrete instrument settings. Carr and Kitagawa (2021) extend this line of work to environments with covariates and propose joint tests, including power-enhancing distillation procedures. Coulibaly et al. (2024), motivated by judge-leniency designs, emphasize the empirical value of exploiting all observable content of the maintained identifying assumptions. Our paper complements these contributions by studying semiparametric MTE specifications with continuous instruments and generated propensity scores, where the observable content of the maintained model combines index sufficiency with stochastic monotonicity and the two components must be handled jointly. Our testing procedures and test

statistics also differ from theirs because we adopt an integrated conditional-moment (ICM) approach that leverages the LCM operator.

Our treatment of the index sufficiency restriction also contributes to the broader literature on specification tests for single-index models. The most closely related paper is Maistre and Patilea (2019), which develops nonparametric model checks for single-index assumptions when the index involves a generated regressor. Their approach relies on kernel smoothing, under which the estimation effect from the first-stage propensity score becomes asymptotically negligible. By contrast, our ICM-based approach, building on Delgado and González-Manteiga (2001) and Escanciano, Jacho-Chávez and Lewbel (2014, EJL14 hereafter), derives influence function representations that explicitly account for these generated regressor and kernel estimation effects. This distinction allows our framework to deliver valid inference under standard root-n estimation rates.

Our paper is also related to the literature that leverages the LCM operator for testing monotonicity conditions; see, e.g., Delgado and Escanciano (2012), Delgado and Escanciano (2013), Seo (2018), Hong and Li (2018, 2020). We adapt these procedures to our context, which involves assessing monotonicity with respect to a generated propensity score.

Organization of the paper: The rest of the paper is organized as follows. Section 2 presents the model setup and the testable implications. Section 3 discusses the construction of test statistics. Section 4 presents the simulation results. Section 5 provides an empirical application. Section 6 concludes. All proofs are presented in the appendix.

2 Model and Sharp Testable Implications

2.1 Model Setting

Our data sample consists of $\{Y, D, X, Z_0\}$, where Y is the observed outcome taking values from the support $\mathcal{X}_Y \subseteq \mathbf{R}$, $D \in \{0, 1\}$ denotes the observed treatment indicator (e.g., college attendance), Z_0 is the vector of IVs taking values in $\mathcal{X}_{Z_0} \subseteq \mathbf{R}^{d_{z_0}}$, and X is the vector of covariates taking values in $\mathcal{X}_X \subseteq \mathbf{R}^{d_x}$. Henceforth, $\mathcal{X}_\xi \subseteq \mathbf{R}^{d_\xi}$ denotes the support of the generic random variable ξ with dimension d_ξ. Henceforth, we write $Z = (X, Z_0)$.

Let $Y(d, z_0)$ denote the potential outcome if an individual receives treatment $D = d$ and the instrumental variables were set to $Z_0 = z_0$. The observed outcome Y is then $Y = DY(1, Z_0) + (1 - D)Y(0, Z_0)$. This general formulation allows for potential direct effects of Z_0 on outcomes, which we will restrict in our assumptions.

In our setting, Z_0 includes at least one continuous component, as the local IV identification strategy for MTEs typically requires. The treatment assignment mechanism is conceptualized as a Roy selection model, in which treatment D is determined by a latent index that includes the observed components Z and an unobserved component U_D.

The following three assumptions constitute the core identifying restrictions of the MTE model.

Assumption 1 (Instrument Exclusion Restriction)**.** *For each $d \in \{0,1\}$, the potential outcome does not depend on the instrument, in the sense that*

$$Y(d, z_0) = Y(d, z_0') \equiv Y(d) \quad a.s.\ for\ all\ z_0, z_0' \in \mathcal{X}_{Z_0},$$

where $Y(d)$ denotes the common potential outcome under treatment state d.

Assumption 2 (Quasi-Random Assignment)**.** *For $d \in \{0,1\}$,*

$$(Y(d, z_0),\, U_D; z_0 \in \mathcal{X}_{Z_0}) \perp Z_0 \mid X.$$

Assumption 3 (Single Threshold-Crossing)**.** *The latent index model for treatment assignment mechanism is governed by the threshold crossing model $D = \mathbf{1}\{\nu(Z) - U_D \geqslant 0\}$ for a measurable and nontrivial unknown function $\nu(\cdot)$, where U_D is a scalar unobservable variable with a continuous distribution.*

Assumption 1 rules out any direct causal effect of Z_0 on outcomes, allowing potential outcomes to be indexed solely by treatment status d. Assumption 2 requires that, conditional on X, the instruments are assigned independently of both potential outcomes and the unobserved resistance to treatment U_D. This is the relevant exogeneity condition in the MTE framework. Assumption 3, as shown by Vytlacil (2002), is equivalent to the IV monotonicity restriction of Imbens and Angrist (1994). Economically, it requires all units to respond to instrument variation in a common direction, ruling out defiers.

Under Assumptions 1–3, the MTE model can be expressed as follows:

$$\begin{aligned} Y &= DY(1) + (1-D)Y(0), \\ D &= \mathbf{1}\{\nu(Z) - U_D \geqslant 0\}. \end{aligned} \tag{2.1}$$

In the model, $\nu(Z)$ captures the observed component of the latent utility from treatment, while U_D represents the unobserved resistance or disutility, which lowers the likelihood of participation.

Because Assumption 2 requires independence conditional on X, the normalization of the resistance to treatment must also be performed conditional on each value $X = x$. Let $U_D^* = F_{U_D|X}(U_D \mid x)$, so that $U_D^* \mid X = x$ is uniformly distributed on $[0,1]$. The treatment rule becomes $D = \mathbf{1}\{P(Z) \geqslant U_D^*\}$ with propensity score $P(Z) = F_{U_D|X}(\nu(Z) \mid X) = \Pr(D = 1 \mid Z)$. For notational simplicity, we henceforth write P for $P(Z)$ and U_D for U_D^*.

Under the threshold-crossing representation, the latent variable U_D indexes individuals by their treatment resistance. Individuals with lower values of U_D are more likely to select into treatment for a given value of Z. This motivates the marginal treatment effect (MTE), defined as

$$\mathrm{MTE}(x, u) = E[Y(1) - Y(0) \mid X = x, U_D = u]\,. \tag{2.2}$$

Heckman and Vytlacil (1999) show that, under Assumptions 1–3 and when the propensity score varies continuously, the MTE curve is nonparametrically identified on the support of P through the local instrumental variables (LIV) estimand:

$$\mathrm{MTE}(x,u) = \frac{\partial E(Y \mid X = x, P = u)}{\partial u}, \qquad u \in \mathcal{X}_P. \tag{2.3}$$

2.2 Testable Implications

Though Assumptions 1–3 involve the latent variables $(U_D, Y(1), Y(0), Z)$ and are therefore not directly testable, they impose restrictions on the joint distribution of the observed variables (Y, D, Z). These restrictions constitute the testable implications of Assumptions 1–3. We restate the testable implications proposed by Heckman and Vytlacil (2005) and establish their sharpness in Proposition 2.1.

Proposition 2.1. *Let (Y, D, Z) and the potential outcomes $\{Y(d, z_0)\}_{d\in\{0,1\}, z_0\in\mathcal{X}_{Z_0}}$ satisfy the potential outcome model $Y = DY(1, Z_0) + (1 - D)Y(0, Z_0)$. Suppose Assumptions 1–3 hold. Define the propensity score as $P = P(Z) = \Pr(D = 1 \mid Z)$. Then the following statements hold.*
(i) Index sufficiency. For every $y \in \mathcal{X}_Y$, each $d \in \{0, 1\}$, and any $z = (x, z_0) \in \mathcal{X}_Z$ such that $p = P(z)$,

$$\Pr(Y \leqslant y \mid P = p, X = x, D = d) = \Pr(Y \leqslant y \mid Z = z, D = d). \tag{2.4}$$

(ii) Nesting inequalities. For every $y_1, y_2 \in \mathcal{X}_Y$ with $y_1 \leqslant y_2$, any pair $0 < p' \leqslant p < 1$, and every $x \in \mathcal{X}_X$,

$$\begin{aligned} &\Pr(y_1 \leqslant Y \leqslant y_2, D = 1 \mid P = p, X = x) \geqslant \Pr(y_1 \leqslant Y \leqslant y_2, D = 1 \mid P = p', X = x), \\ &\Pr(y_1 \leqslant Y \leqslant y_2, D = 0 \mid P = p, X = x) \leqslant \Pr(y_1 \leqslant Y \leqslant y_2, D = 0 \mid P = p', X = x). \end{aligned} \tag{2.5}$$

(iii) Sharp characterization. Conversely, if the observable distribution of (Y, D, Z) satisfies (2.4) and (2.5) for the given propensity score $P(Z)$, then there exist a scalar latent variable $\tilde{U}_D$, satisfying $\tilde{U}_D \mid X = x \sim U[0, 1]$ for all $x \in \mathcal{X}_X$ and $\tilde{U}_D \perp Z_0 \mid X$, a treatment rule $\tilde{D} = \mathbf{1}\{P(Z) \geqslant \tilde{U}_D\}$, and potential outcomes $\{\tilde{Y}(d)\}_{d\in\{0,1\}}$ such that $(\tilde{Y}(1), \tilde{Y}(0), \tilde{U}_D) \perp Z_0 \mid X$ and $\tilde{Y}(d, z_0) = \tilde{Y}(d)$ for all $z_0 \in \mathcal{X}_{Z_0}$. The constructed model satisfies Assumptions 1–3, and $(\tilde{Y}, \tilde{D}, Z)$ has the same distribution as (Y, D, Z).

Parts (i) and (ii) state the full set of testable implications proposed by Heckman and Vytlacil (2005): the single index sufficiency condition in (2.4) and the nesting inequalities in (2.5). As noted by Carr and Kitagawa (2021), these two restrictions play complementary roles and should ideally be assessed jointly. Relative to the class of all nonnegative functions of Y considered by Heckman and Vytlacil (2005), Proposition 2.1 specializes the relevant test functions separately for the two observable implications: the indicators $\mathbf{1}(Y \leqslant y)$ characterize the conditional distribution function

and therefore fully capture the equality restrictions implied by index sufficiency, whereas the interval indicators $\mathbf{1}(y_1 \leqslant Y \leqslant y_2)$ characterize interval probabilities and therefore fully capture the stochastic ordering restrictions embodied in the nesting inequalities. The proposition, therefore, replaces an unwieldy class of test functions with a tractable subclass that preserves the full informational content of the observable restrictions. Part (iii) then establishes that (2.4) and (2.5) provide a sharp characterization of Assumptions 1–3.

Related sharpness results for the nesting inequalities have been established in settings where the propensity score is treated as primitive; see, for example, Mourifié and Wan (2025); Coulibaly et al. (2024). We build on this line of work by explicitly accounting for covariates X while establishing sharpness. In addition, our formulations start from the empirically observed vector (Y, D, Z) rather than (Y, D, P) (or (Y, D, X, P)). Maintaining a propensity score specification $z \mapsto P(z)$ allows the nesting inequalities to be stated in terms of the scalar $P(Z)$. At the (Y, D, Z) level, however, sharpness also requires that $P(Z)$ exhaust the outcome relevant information in Z; this additional requirement is precisely the index sufficiency restriction. In this sense, Proposition 2.1 provides the sharpness characterization of the MTE assumptions based on all observable variables (Y, D, Z).

While Proposition 2.1 provides a useful benchmark characterization conditional on covariates, implementing a fully nonparametric test based on these restrictions becomes difficult in empirical applications when X is of moderate dimensions. In that case, both testing the observable restrictions and nonparametric identification of the MTE curve are hindered by limited common support in the propensity score across treated and untreated units.

To obtain an operational framework that remains close to applied practice, we therefore adopt a semiparametric specification that treats covariates parametrically, following Carneiro et al. (2011); Schoenberg, Cornelissen, Dustmann and Raute (2018) and, more recently, Carr and Kitagawa (2021). Throughout this semiparametric specification, we take X to include a constant term and introduce two additional assumptions:

Assumption 4 (Linear Functional Form)**.** *For $d \in \{0, 1\}$ and $z_0 \in \mathcal{X}_{Z_0}$, the potential outcomes take the linear form $Y(d, z_0) = X^\top \beta_d + V(d, z_0)$.*

Assumption 5 (Full Independence)**.** *For $d \in \{0, 1\}$,*

$$(V(d, z_0),\, U_D; z_0 \in \mathcal{X}_{Z_0}) \perp (Z_0, X).$$

Assumptions 4 and 5 are standard in empirical applications and substantially relax the support requirements for identifying the MTE curve. In particular, Carneiro et al. (2011) show that, under these two additional assumptions, the conditional expectation of the outcome admits the following partial linear representation:

$$E(Y|X = x, P = p) = x^\top \beta_0 + p x^\top (\beta_1 - \beta_0) + m(p), \tag{2.6}$$

where $m(\cdot)$ is an unknown function of the propensity score. It then follows that the MTE curve is identified as

$$\mathrm{MTE}(x,u) = x^\top(\beta_1 - \beta_0) + m'(u). \tag{2.7}$$

This partial linear structure implies that the unconditional full support of the propensity score on the unit interval is sufficient to identify the entire MTE curve. This requirement is substantially weaker than the conditional full support condition needed in the absence of Assumptions 4 and 5. Without these additional assumptions, identification through the LIV estimand in (2.3) requires that the propensity score have full support on the unit interval conditional on the covariates for both the treated and untreated groups. Consequently, the semiparametric approach is arguably more realistic in many empirical applications.

When Assumptions 4 and 5 hold, the testable implications of (2.4) and (2.5) are recharacterized in terms of the distribution of $(V(1,z_0), V(0,z_0))$ rather than $(Y(1,z_0), Y(0,z_0))$[1].

Proposition 2.2. *Let (Y,D,Z) and the potential outcomes $\{Y(d,z_0)\}_{d\in\{0,1\},z_0\in\mathcal{X}_{Z_0}}$ satisfy $Y = DY(1,Z_0) + (1-D)Y(0,Z_0)$. Suppose Assumptions 1, 3, 4, and 5 hold. Under Assumptions 1 and 4, write*

$$Y(d) = X^\top\beta_d + V(d), \qquad d\in\{0,1\},$$

where $V(d) = Y(d) - X^\top\beta_d$. Let

$$V = DV(1) + (1-D)V(0) = D\{Y - X^\top\beta_1\} + (1-D)\{Y - X^\top\beta_0\},$$

and let $P = P(Z) = \Pr(D=1 \mid Z)$. Then the following statements hold.
(i) Index sufficiency. For every $v\in\mathcal{X}_V$, each $d\in\{0,1\}$, and any $z\in\mathcal{X}_Z$ such that $p = P(z)$,

$$\Pr(V \leqslant v \mid P=p, D=d) = \Pr(V\leqslant v \mid Z=z, D=d). \tag{2.8}$$

(ii) Nesting inequalities. For every $v_1 \leqslant v_2$ and any pair $0 < p' \leqslant p < 1$,

$$\begin{aligned} \Pr(v_1 \leqslant V \leqslant v_2, D=1 \mid P=p) &\geqslant \Pr(v_1 \leqslant V \leqslant v_2, D=1 \mid P=p'),\\ \Pr(v_1 \leqslant V \leqslant v_2, D=0 \mid P=p) &\leqslant \Pr(v_1 \leqslant V \leqslant v_2, D=0 \mid P=p'). \end{aligned} \tag{2.9}$$

(iii) Sharp characterization under the maintained semiparametric structure. Conversely, fix (β_0,β_1) and form the observed residual V above. If the residual distribution of (V,D,Z) satisfies (2.8) and (2.9) for the given propensity score $P(Z)$, then there exist a scalar $\tilde{U}_D \sim U[0,1]$, a treatment rule $\tilde{D} = \mathbf{1}\{P(Z) \geqslant \tilde{U}_D\}$, and latent residuals $\{\tilde{V}(d)\}_{d\in\{0,1\}}$ such that

$$(\tilde{V}(d), \tilde{U}_D) \perp Z, \qquad \tilde{Y}(d,z_0) = X^\top\beta_d + \tilde{V}(d), \qquad d\in\{0,1\},\ z_0\in\mathcal{X}_{Z_0}.$$

[1]To be precise, the residuals derived from (2.6) are $\hat{V}(1,z_0) = V(1,z_0) + c_1$ and $\hat{V}(0,z_0) = V(0,z_0) + c_2$ for two constants c_1 and c_2. A location shift in relation to $(V(1,z_0), V(0,z_0))$ does not influence the testable implications.

The resulting model satisfies Assumptions 1, 3, 4, and 5, and $(\tilde{Y}, \tilde{D}, Z)$ *has the same distribution as* (Y, D, Z).

Remark 2.1 (Oracle Interpretation of Sharpness)**.** Proposition 2.2 establishes sharpness conditional on the maintained semiparametric structure: the linear outcome specification in Assumption 4 used to form V and the propensity score specification $z \mapsto P(z)$, both treated as given. In this sense, the result is an *oracle* sharpness statement: the residual V and the propensity score P are taken as known objects. Their estimation effects enter the testing problem in Section 3 via influence function corrections, rather than through the sharpness characterization itself.

The semiparametric characterization in Proposition 2.2 is useful not only because it yields restrictions that remain operational in empirical settings with many covariates, but also because it clarifies the sense in which the residual-based implications are informative about IV identifying assumptions. In particular, the restrictions in (2.8) and (2.9) are sharp only in the oracle sense described in Remark 2.1. The empirical implementation in Section 3 additionally requires a correctly specified parametric propensity score model, since the feasible test replaces the oracle propensity score by its first step estimate. By contrast, if either maintained component fails, the same restrictions may still be viewed as testable implications, but they no longer admit the sharp interpretation given in Proposition 2.2. As Carr and Kitagawa (2021) note, a rejection in this setting can therefore reflect either a violation of these IV identifying assumptions or a failure of the maintained parametric components. However, there is a large body of specification testing literature that can be used to assess the goodness of fit of these nuisance models; see, e.g., Sant'Anna and Song (2019) for propensity score tests.

This formulation also helps clarify the relation between our procedure and the existing literature. Carr and Kitagawa (2021) study the same broad semiparametric implications and likewise emphasize the complementarity between index sufficiency and nesting inequalities. Although their framework accommodates continuous instruments, their implementation handles that case by discretizing the instrument after residualization and then using distillation to sharpen the monotonicity component through instrument-indexed subdensities. Coulibaly et al. (2024), by contrast, retain the continuous propensity score as the central object but focus on monotonicity alone, reformulating (2.9) as unconditional moment inequalities. Our approach differs from both in that it jointly evaluates the original sharp semiparametric implications and retains the continuous propensity score structure throughout, thereby keeping the procedure aligned with the full observable content of IV identifying assumptions in the maintained semiparametric MTE model. As a direct consequence, our testing procedures significantly differ from those currently available, though we stress that they should be viewed as complementary. We provide more details about this in the next session.

3 Test Statistics

In this section, we focus on constructing falsification tests for the MTE assumptions based on Proposition 2.2, as we view this setting as more empirically relevant.[2] We first rewrite index sufficiency in (2.8) and the nesting inequalities in (2.9) as operational null hypotheses in terms of observable objects, and then construct the associated marginal and joint test statistics. A central econometric difficulty is that both the propensity score and the outcome residuals are generated regressors, so the asymptotic analysis must account for their estimation effects throughout.

3.1 Transformation of Null Hypothesis

We now reformulate each of the two sharp implications in Proposition 2.2 as an operational null hypothesis that can be assessed using sample analogs.

The index sufficiency restriction in (2.8) requires that, within the subpopulation with treatment status $D = d$, the conditional distribution of the residual V depends on the instrument vector Z only through the propensity score P. When this holds, the conditional distribution of V given $P, Z, D = d$ collapses to the conditional distribution of V given $P, D = d$. So any residual covariation between V and Z beyond what is captured by P constitutes a violation. This places our problem in the literature on testing conditional independence; see Delgado and González-Manteiga (2001) and Escanciano et al. (2014).

We operationalize this observation through an integrated conditional moment (ICM) criterion. For $d \in \{0, 1\}$ and $\omega = (v, p, z) \in \mathscr{W} \equiv \mathcal{X}_V \times \mathcal{X}_P \times \mathcal{X}_Z$, define

$$U_d(\omega) \equiv E\big[\big(\mathbf{1}(V \leqslant v) - \Pr(V \leqslant v \mid P, D = d)\big)\mathbf{1}(P \leqslant p, Z \leqslant z, D = d)\big]. \tag{3.1}$$

The function $U_d(\omega)$ captures whether the centered residual indicator $\mathbf{1}(V \leqslant v) - \Pr(V \leqslant v \mid P, D = d)$ still covaries with the instrument indexed weight $\mathbf{1}(P \leqslant p, Z \leqslant z, D = d)$. Under the null, this covariation vanishes for every ω and d. The index sufficiency is therefore equivalent to

$$H_0^I : \ U_d(\omega) = 0, \tag{3.2}$$

for all $\omega \in \mathscr{W}$ and $d \in \{0, 1\}$. A key practical advantage of this representation is that Z enters (3.1) only as an argument of the indicator weight $\mathbf{1}(P \leqslant p, Z \leqslant z, D = d)$, not as a conditioning variable. Consequently, the test requires nonparametric smoothing only over the scalar propensity score P, and the potentially high-dimensional instrument vector Z is handled without any additional smoothing.

[2] We conjecture that the techniques we use in the paper can be applied to formulating tests based on Proposition 2.1, provided we adopt parametric models for the propensity score. We also anticipate that one may need to adapt the tools discussed in Section 3.2 of Delgado and Escanciano (2013) to account for d-dimensional X. From a practical standpoint, as noted above, support restrictions in such settings can be taxing across many applications. We leave a discussion of these extensions to future work.

An additional complication is that both P and V are generated regressors, so their estimation effects must be properly accounted for in the limiting distribution of the test statistic. We address this through new influence function representations, developed in Appendix B.

We next consider the nesting inequalities in (2.9). These restrictions impose monotonicity on the outcome-treatment subdistributions. For each interval $[v_1, v_2]$, the conditional probability mass $\Pr(v_1 \leqslant V \leqslant v_2, D = 1 \mid P = p)$ must be nondecreasing in p, whereas $\Pr(v_1 \leqslant V \leqslant v_2, D = 0 \mid P = p)$ must be nonincreasing in p. Testing this family of ordering restrictions over a continuous argument p is not straightforward: pointwise comparisons accumulate over an uncountable index set, and a direct supremum approach requires choosing how to discretize the propensity score, at the cost of information loss.

To avoid this, we follow Delgado and Escanciano (2012) and Seo (2018) and integrate the conditional subdistribution against the rank of P. The key insight is that the ordering restrictions in (2.9) are equivalent, under continuity of P, to a concavity requirement on a copula-type functional that depends only on the quantile rank u of the propensity score rather than its level. Formally, for $d \in \{0, 1\}$, $v_1 \leqslant v_2$, and $u \in [0, 1]$, define

$$\begin{aligned} C_d(v_1, v_2, u) &= \int_0^u \Pr(v_1 \leqslant V \leqslant v_2, D = d \mid P = F_P^{-1}(\bar{u}))\, d\bar{u} \\ &= \Pr(v_1 \leqslant V \leqslant v_2, D = d, P \leqslant F_P^{-1}(u)) \\ &\equiv H_d(v_1, v_2, F_P^{-1}(u)), \end{aligned} \tag{3.3}$$

where $F_P^{-1}(\cdot)$ is the quantile function of P. The function $C_d(v_1, v_2, u)$ accumulates the conditional subdistribution probability over propensity score quantiles up to rank u; its slope in u is therefore the conditional probability mass $\Pr(v_1 \leqslant V \leqslant v_2, D = d | P = F_P^{-1}(u))$.

By Nelsen (2007), under continuity of P, the nesting inequalities in (2.9) hold if and only if $C_0(v_1, v_2, \cdot)$ is concave in u and $C_1(v_1, v_2, \cdot)$ is convex in u, for each fixed $v_1 \leqslant v_2$. This concavity characterization, in turn, admits a reformulation as an equality restriction that is amenable to a root-n test statistic. Let $\tilde{\theta}$ denote the partial least concave majorant (LCM) operator, which extends the standard LCM to accommodate the joint dependence on $\lambda = (v_1, v_2, u)$; Appendix B provides its formal definition. Then the monotonicity null is

$$H_0^M : \mathcal{T}_0 C_0(\lambda) \equiv \tilde{\theta} C_0(\lambda) - C_0(\lambda) = 0 \quad \text{and} \quad \mathcal{T}_1 C_1(\lambda) \equiv \tilde{\theta}(-C_1)(\lambda) + C_1(\lambda) = 0, \tag{3.4}$$

Both operators measure the gap between a function and its least concave majorant in u, with $\mathcal{T}_0$ acting on C_0 and $\mathcal{T}_1$ on $-C_1$. Under H_0^M, both target functions are already concave in u, so the gaps are identically zero; under a violation, the least concave majorant strictly dominates at some λ, and the corresponding gap becomes positive.

At points where the nesting inequalities bind, $C_d(v_1, v_2, \cdot)$ is locally linear in u, reflecting that the

concavity restriction is satisfied at the boundary rather than strictly. We collect such points in the contact set B_C. The behavior of B_C is critical for size control: when B_C has positive measure, the test statistic has a nondegenerate limit under H_0^M, and the bootstrap delivers asymptotically exact size; when B_C has measure zero, the limit distribution degenerates and the test becomes conservative. While the nesting inequalities can also be approached through the generalized regression monotonicity framework of Hsu, Liu and Shi (2019), as in Coulibaly et al. (2024), our formulation is tailored to joint testing with index sufficiency and provides correct size control when the contact set has positive measure. We view these approaches as complements.

3.2 Testing Procedures

We now construct feasible sample analogs of the null hypotheses in (3.2) and (3.4), and use them to define the marginal and joint test statistics. The main implementation issue is that both the residuals and the propensity score are unobserved and must be estimated in a first step. As a result, the asymptotic null distribution of the test statistics must account for the generated regressor nature of $\hat{V}_i$ and $\hat{P}_i$. Our inference procedure addresses this issue using a multiplier bootstrap based on the corresponding influence function representations. The resulting procedure is summarized as follows:

Algorithm.

i. Estimate the propensity score P_i under the postulated parametric specification to obtain $\hat{P}_i$, and estimate the partially linear model (2.6) to obtain the treatment specific residuals $\hat{V}_{di} = Y_i - X_i^\top \hat{\beta}_d$ for $d \in \{0, 1\}$.

ii. Compute the marginal test statistics:

$$\hat{T}_I = \sqrt{n} \max_{d\in\{0,1\}} \left\| \hat{U}_d(\omega) \right\|_\infty, \qquad \hat{T}_M = \sqrt{n} \max_{d\in\{0,1\}} \left\| \mathcal{T}_d \hat{C}_d(\lambda) \right\|_\infty. \tag{3.5}$$

and the joint statistics for $H_0^I \cap H_0^M$:

$$\hat{T}_{\text{Max}} = \max\left\{ \hat{T}_I, \hat{T}_M \right\}, \qquad \hat{T}_{\text{Sum}} = \hat{T}_I + \hat{T}_M. \tag{3.6}$$

iii. For $b = 1, \ldots, B$, implement the multiplier bootstrap:

(a) Generate multiplier weights $\{\xi_i\}_{i=1}^n$ that are bounded, mean zero, variance one, and independent of the sample. Within each bootstrap iteration, the same multiplier draw is used for both the index-sufficiency and monotonicity bootstrap components defined below.

(b) Construct the bootstrap empirical processes:

$$\begin{aligned}\hat{U}_d^\star(\omega) &= \frac{1}{n}\sum_{i=1}^n \xi_i\, \hat{\zeta}_d^I\left(Y_i, D_i, Z_i; \omega, \hat{\alpha}, \hat{\beta}_d\right), \\ \hat{\mathbb{G}}_{C_d}^\star(\lambda) &= \frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i\, \hat{\zeta}_d^M\left(Y_i, D_i, Z_i; \lambda, \hat{\alpha}, \hat{\beta}_d\right),\end{aligned} \tag{3.7}$$

where the estimated influence functions $\hat{\zeta}_d^I$ and $\hat{\zeta}_d^M$ are defined below.

(c) Compute the bootstrap statistics

$$\begin{aligned}\hat{T}_I^\star &= \sqrt{n}\max_{d\in\{0,1\}} \|\hat{U}_d^\star\|_\infty, \\ \hat{T}_M^\star &= \max_{d\in\{0,1\}} \left\|\hat{\mathcal{T}}_{d,\hat{C}_d}\left(\hat{\mathbb{G}}_{C_d}^\star\right)\right\|_\infty,\end{aligned}$$

Then, construct the bootstrap joint statistics:

$$\hat{T}_{\text{Max}}^\star = \max\left\{\hat{T}_I^\star, \hat{T}_M^\star\right\}, \qquad \hat{T}_{\text{Sum}}^\star = \hat{T}_I^\star + \hat{T}_M^\star. \tag{3.8}$$

iv. Set the critical values $\hat{c}_{I,\alpha}^\star$, $\hat{c}_{M,\alpha}^\star$, $\hat{c}_{\text{Max},\alpha}^\star$, and $\hat{c}_{\text{Sum},\alpha}^\star$ equal to the empirical $(1-\alpha)$ quantiles of the B bootstrap replications of $\hat{T}_I^\star$, $\hat{T}_M^\star$, $\hat{T}_{\text{Max}}^\star$, and $\hat{T}_{\text{Sum}}^\star$, respectively. Reject H_0^I when $\hat{T}_I > \hat{c}_{I,\alpha}^\star$ and reject H_0^M when $\hat{T}_M > \hat{c}_{M,\alpha}^\star$. For the joint null $H_0^I \cap H_0^M$, reject when either $\hat{T}_{\text{Max}} > \hat{c}_{\text{Max},\alpha}^\star$ or $\hat{T}_{\text{Sum}} > \hat{c}_{\text{Sum},\alpha}^\star$, depending on the chosen joint statistic. In addition, let $\hat{p}_I$ and $\hat{p}_M$ denote the bootstrap p-values for the two component tests and report the Bonferroni joint p-value

$$\hat{p}_{BC} = \min\{1, 2\min(\hat{p}_I, \hat{p}_M)\},$$

rejecting $H_0^I \cap H_0^M$ at level α when $\hat{p}_{BC} \leqslant \alpha$.

Several aspects of the implementation merit further discussion.

Estimated influence functions. Because the test statistics depend on estimated propensity score $\hat{P}_i$ and the outcome residuals $\hat{V}_{di}$, the bootstrap must account for the additional sampling variability they introduce. The influence functions $\hat{\zeta}_d^I$ and $\hat{\zeta}_d^M$ entering the bootstrap empirical processes (3.7) achieve this correction and take the form:

$$\begin{aligned}\hat{\zeta}_d^I\left(Y_i, D_i, Z_i; \omega, \hat{\alpha}, \hat{\beta}_d\right) = [\mathbf{1}(\hat{V}_{di} \leqslant v) - \hat{F}_{V|P,D}(v \mid \hat{P}_i, d; \hat{\beta}_d, \hat{\alpha})]\hat{\psi}_i^\perp(p, z, d; \hat{\alpha}) \\ + \hat{A}_d^\top \hat{l}\,(D_i, Z_i, \hat{\alpha}) + \hat{B}_d^\top \hat{l}_d(Y_i, Z_i, \hat{\beta}_d, \hat{\alpha}),\end{aligned}$$

and

$$\hat{\zeta}_d^M\left(Y_i, D_i, Z_i; \lambda, \hat{\alpha}, \hat{\beta}_d\right) = \mathbf{1}(v_1 \leqslant \hat{V}_{di} \leqslant v_2,\ D_i = d,\ \hat{F}_P(\hat{P}_i) \leqslant u) - \hat{C}_d(\lambda) - \partial_u \hat{C}_d(\lambda)\{\mathbf{1}(\hat{F}_P(\hat{P}_i) \leqslant u) - u\} + \hat{l}_d^\top(Y_i, Z_i, \hat{\beta}_d, \hat{\alpha})\partial_{\beta_d}\hat{C}_d(\lambda) + \hat{l}^\top(D_i, Z_i, \hat{\alpha})\partial_\alpha \hat{C}_d(\lambda).$$

The first term in each expression is the leading empirical process contribution, while the remaining correction terms account for first-step parametric estimation through the asymptotic linear representations of $\hat{\alpha}$ and $\hat{\beta}_d$. Here, $\hat{l}(D_i, Z_i, \hat{\alpha})$ and $\hat{l}_d(Y_i, Z_i, \hat{\beta}_d, \hat{\alpha})$ are the plug-in influence functions associated with $\hat{\alpha}$ and $\hat{\beta}_d$, respectively (see Assumption A.1(iv)). All terms appearing in these influence functions, along with their detailed derivations, are reported in Appendix B.

Evaluation grids. The sample analogs $\hat{U}_d(\omega)$ and $\hat{C}_d(\lambda)$ correspond to (3.2) and (3.4). Because $\hat{T}_I$ tests the full residual distribution rather than only a conditional mean, its natural sample grid is, for each $d \in \{0, 1\}$, the Cartesian product of residual thresholds and sample induced (p, z) points:

$$\{\hat{V}_{di} : D_i = d\} \times \{(\hat{P}_j, Z_j) : j = 1, \ldots, n\}.$$

Similarly, $\hat{T}_M$ is naturally evaluated on the Cartesian product, for each $d \in \{0, 1\}$,

$$\{[\hat{V}_{di}, \hat{V}_{dj}] : \hat{V}_{di} \leqslant \hat{V}_{dj},\ D_i = D_j = d\} \times \{0, 1/n, \ldots, 1\},$$

To reduce computation, we approximate only the residual dimension by a quantile grid $\{v_k : k = 1, \ldots, M\}$. Hence $\hat{T}_I$ is evaluated on $\{v_k\}_{k=1}^M \times \{(\hat{P}_i, Z_i)\}_{i=1}^n$, while $\hat{T}_M$ is evaluated on $\{[v_k, v_\ell] : v_k \leqslant v_\ell,\ k, \ell = 1, \ldots, M\} \times \{0, 1/n, \ldots, 1\}$. The (p, z) coordinates in $\hat{T}_I$ and the rank coordinate in $\hat{T}_M$ are evaluated on their sample-induced grids.

Bandwidth selection. The bandwidth h is the only smoothing parameter in the test statistics, serving two purposes: the nonparametric partialling-out step in Robinson's partially linear estimator for (2.6), and the nonparametric estimation of $\hat{F}_{V|P,D}$ entering the index statistic $\hat{T}_I$ in (3.5). Additional smoothing parameters arise only in auxiliary calculations for the bootstrap correction terms $\hat{B}_d$ and $\partial_u \hat{C}_d$. These do not affect the leading test statistics, and we report their specific values in Section 4.

Numerical delta method. Implementation of the monotonicity bootstrap requires the directional derivative of the gap operator $\mathcal{T}_d$ in (3.4). Let $\hat{\dot{\mathcal{T}}}_d$ denote the numerical approximation to this directional derivative, evaluated at the estimated function $\hat{C}_d$ and applied to a perturbation direction.

Following Hong and Li (2018, 2020), for the bootstrap direction $\hat{\mathbb{G}}^{\star}_{C_d}$, we define

$$\hat{\dot{\mathcal{T}}}_d\Big(\hat{\mathbb{G}}^{\star}_{C_d}\Big) = \frac{\mathcal{T}_d\Big(\hat{C}_d + \kappa\hat{\mathbb{G}}^{\star}_{C_d}\Big) - \mathcal{T}_d\Big(\hat{C}_d\Big)}{\kappa}, \tag{3.9}$$

where $\kappa = \kappa_n$ satisfies $\kappa \to 0$ and $\sqrt{n}\kappa \to \infty$. This numerical delta method requires only two evaluations of $\mathcal{T}_d$, avoids explicit estimation of the contact set, and satisfies the conditions of Lemma S.3.8 in Fang and Santos (2018) for consistent estimation of the Hadamard directional derivative of the LCM operator.

Remark 3.1 (Joint testing vs. Bonferroni correction)**.** We view the joint procedures as complementary summaries of the two observable components. The component p-values are substantively important and should always be reported. The max and sum statistics provide omnibus joint tests based on the joint multiplier process, while the Bonferroni combination offers a transparent scale robust benchmark. In finite samples, the unweighted max and sum statistics may be affected by scale differences between the index and monotonicity components, so the Bonferroni procedure can be more informative in some designs.

Theorem 3.1 establishes that the bootstrap procedure controls size under the null and is consistent under fixed alternatives. The proof is given in Appendix D. The regularity conditions required for the asymptotic validity are collected in Appendix A. Throughout the theorem, probabilities are taken with respect to the original sampling distribution. The bootstrap critical values are the conditional $(1-\alpha)$ quantiles of the corresponding bootstrap statistics given the original sample.

Theorem 3.1 (Bootstrap validity for the observable tests)**.** *Suppose Assumptions A.1–A.6 in Appendix A hold. Let*

$$\Delta_I = \max_{d\in\{0,1\}} \|U_d\|_\infty, \qquad \Delta_M = \max_{d\in\{0,1\}} \|\mathcal{T}_d C_d\|_\infty.$$

A fixed alternative to H_0^I means $\Delta_I > 0$, and a fixed alternative to H_0^M means $\Delta_M > 0$.

(i) *Under H_0^I,*

$$\lim_{n\to\infty} \Pr(\hat{T}_I \geqslant \hat{c}^{\star}_{I,\alpha}) = \alpha.$$

Under fixed alternatives to H_0^I,

$$\lim_{n\to\infty} \Pr(\hat{T}_I \geqslant \hat{c}^{\star}_{I,\alpha}) = 1.$$

(ii) *Under H_0^M, if the contact set has a positive measure, then*

$$\lim_{n\to\infty} \Pr(\hat{T}_M \geqslant \hat{c}^{\star}_{M,\alpha}) = \alpha.$$

If the contact set B_C has measure zero, then for any $\eta > 0$, define $\check{T}^{\star}_M = \max\{\eta, \hat{T}^{\star}_M\}$ and let

$\check{c}^{\star}_{M,\alpha}$ denote the regularized bootstrap critical value. This gives a conservative test such that

$$\lim_{n\to\infty} \Pr(\hat{T}_M \geqslant \check{c}^{\star}_{M,\alpha}) = 0.$$

Under fixed alternatives to H_0^M with $\Delta_M > 0$,

$$\lim_{n\to\infty} \Pr(\hat{T}_M \geqslant \hat{c}^{\star}_{M,\alpha}) = 1.$$

(iii) *Under $H_0^I \cap H_0^M$, when the monotonicity component is evaluated in the nondegenerate contact set case, the max and sum joint tests are based on the joint multiplier draw satisfying*

$$\lim_{n\to\infty} \Pr(\hat{T}_{\text{Max}} \geqslant \hat{c}^{\star}_{\text{Max},\alpha}) \leqslant \alpha, \qquad \lim_{n\to\infty} \Pr(\hat{T}_{\text{Sum}} \geqslant \hat{c}^{\star}_{\text{Sum},\alpha}) \leqslant \alpha.$$

The Bonferroni procedure satisfies

$$\lim_{n\to\infty} \Pr(\hat{p}_{BC} \leqslant \alpha) \leqslant \alpha.$$

Under fixed alternatives with $\max\{\Delta_I, \Delta_M\} > 0$, all three joint procedures are consistent, provided the component that detects the violation is combined using its corresponding valid critical value.

Theorem 3.1 shows that the monotonicity test is asymptotically exact when the contact set has positive measure, whereas it becomes conservative when the contact set has measure zero because the limit distribution degenerates. In the measure zero contact set case, the formal conservative statement in part (ii) is established for the regularized bootstrap critical value $\check{c}^{\star}_{M,\alpha}$ based on $\check{T}^{\star}_M = \max\{\eta, \hat{T}^{\star}_M\}$, where the constant $\eta > 0$ plays a purely theoretical role by preventing the bootstrap critical value from collapsing to zero. A similar regularization device appears in Beare and Shi (2018). The same conservative interpretation extends to joint procedures in the degenerate contact set case only when the monotonicity component is combined using the regularized critical value $\check{c}^{\star}_{M,\alpha}$ or its corresponding regularized bootstrap p-value.

In the simulations and empirical application, however, we report the unregularized critical value $\hat{c}^{\star}_{M,\alpha}$ as the baseline implementation. This choice is motivated by the observed finite-sample conservatism of the unregularized procedure in the measure-zero designs N1–N3, where further regularization would only shift the critical value upward, thereby inducing additional power loss. Thus, the formal theorem and the reported baseline coincide in the nondegenerate contact set case; in the degenerate case, the theorem supplies a conservative regularized benchmark, and the unregularized implementation serves as the practical baseline.

4 Simulations

This section studies the finite-sample performance of the proposed bootstrap procedures in terms of both size and power. We report results for the marginal tests $\hat{T}_I$ and $\hat{T}_M$, the joint tests $\hat{T}_{\text{Max}}$ and $\hat{T}_{\text{Sum}}$, and the Bonferroni-corrected combination test $\hat{T}_{\text{BC}}$. The main text focuses on rejection rates at the 5% level for sample sizes $n = 200, 500, 1000, 2000$, while the corresponding results at the 1%, 5%, and 10% levels are reported in Appendix F. Critical values are computed using 300 bootstrap replications, and each design is repeated 1000 times.[3]

We now specify the tuning parameters used in the simulation study, following the implementation described in Section 3. We set $h = 0.5n^{-1/3}$ for the index-sufficiency process $U_d(\omega)$ in (3.1) and $\kappa = 0.15\log(n)n^{-1/2}$ for the numerical delta bootstrap in (3.9). To compute the feasible supremum, we evaluate $\hat{T}_M$ on a percentile-based residual interval grid (2nd to 98th percentiles in increments of 2). The same residual evaluation points are used for $\hat{T}_I$.

The auxiliary bandwidth for estimating $\hat{B}_d$ is set to $\ell = n^{-1/3}$, except in the multimodal designs ALT4(I) and ALT4(M), where we use $\ell = 0.5n^{-1/3}$ to stabilize the finite-sample behavior of $\hat{T}_I$; see the notes to Tables 2– 3. Finally, we use bandwidth $n^{-1/2}$ for $\partial_u \hat{C}_d(\lambda)$ in the auxiliary terms of the influence function. Details on the construction of $\hat{\zeta}_d^I$ and $\hat{\zeta}_d^M$ are provided in Appendix B.

Size Performance

To examine finite sample size, we begin with a Roy selection design under which both the index sufficiency and monotonicity hold. The treatment equation and potential outcomes are given by

$$\begin{aligned} D &= \mathbf{1}(\phi_1 Z_1 + \phi_2 Z_2 + \gamma X + U_D \geqslant 0), \\ Y(0) &= X\beta_0 + V(0), \\ Y(1) &= X\beta_1 + V(1), \\ (V(1), V(0), U_D) &\sim N(0, \Sigma). \end{aligned} \tag{4.1}$$

Here, (X, Z_1, Z_2) are mutually independent standard normal random variables, and $(\phi_1, \phi_2, \gamma, \beta_1, \beta_0) = (1, 1, 1, 2, 1)$. The covariance matrix Σ has ones on the diagonal and a common off-diagonal element ρ.

We consider four null designs, denoted N1–N4, to study how the finite-sample behavior of the tests varies with the joint distribution of $(V(1), V(0), U_D)$. In N1–N3, we set $\rho \in \{0, 0.25, 0.81\}$, corresponding to independence, moderate dependence, and strong dependence, respectively, while maintaining unit marginal variances. Design N4 is a boundary case of this Gaussian family: it specifies a degenerate jointly normal distribution by setting $(V(1), V(0), U_D)^\top = (1, 2, 1)^\top U_D$, where $U_D \sim N(0, 1)$. This design generates a contact set B_C of positive measure, under which the monotonicity statistic $\hat{T}_M$ has a nondegenerate null limit. At the same time, because $\text{Var}(V(0)) = 4$,

[3]We use 300 bootstrap replications in the simulation study because evaluating both components on a high dimensional grid is computationally demanding. The empirical application uses 1000 multiplier replications.

comparisons between N4 and N1–N3 also reflect a larger residual scale for the untreated potential outcome.

Table 1 reports rejection rates at the 5% level. Across N1–N4, the index statistic $\hat{T}_I$ remains close to the nominal level, with rejection rates between 0.035 and 0.059. By contrast, $\hat{T}_M$ is conservative in N1–N3, where the contact set is of measure zero, with rejection rates between 0.002 and 0.016. The low rejection rates for $\hat{T}_M$ in N1–N3 are consistent with the degeneracy arising from a zero contact set, as characterized in Theorem 3.1(ii). This theorem formally establishes a conservative size result when using the regularized critical value $\check{c}^{\star}_{M,\alpha}$. Since the unregularized baseline $\hat{c}^{\star}_{M,\alpha}$ is already empirically conservative in these designs, additional regularization is unnecessary. A similar finite sample pattern has been documented for LCM-based tests; see Beare and Shi (2018) and Fang (2019).

This conservativeness carries over to all the joint statistics, whose rejection frequencies mostly fall between 0.002 and 0.039 in those designs. But these joint statistics exhibit different finite-sample patterns. The rejection rates of $\hat{T}_{\text{Max}}$ increase with the sample size and move toward the nominal level, especially in N1 and N2. For example, in N1 they rise from 0.002 at $n = 200$ to 0.039 at $n = 2000$, and in N2 from 0.004 to 0.027. In contrast, $\hat{T}_{\text{Sum}}$ remains more conservative and shows little improvement as n increases. The Bonferroni statistic $\hat{T}_{\text{BC}}$ is comparatively stable across sample sizes and DGPs, and its rejection rates are numerically close to the average of the rejection rates of $\hat{T}_I$ and $\hat{T}_M$. Accordingly, under measure zero contact sets, $\hat{T}_{\text{BC}}$ appears more reliable than the unweighted joint statistics, which is consistent with the discussion in Remark 3.1.

Design N4 illustrates the distinct behavior that arises when the contact set has positive measure. In this case, the rejection rate of $\hat{T}_M$ increases from 0.011 at $n = 200$ to 0.050 at $n = 2000$, while $\hat{T}_{\text{Sum}}$, $\hat{T}_{\text{Max}}$ and $\hat{T}_{\text{BC}}$ also move closer to the nominal level. Because convergence is gradual in this design, presumably due to the numerical delta bootstrap, we also examined the case $n = 5000$, which is omitted from Table 1. The corresponding rejection rates of $\hat{T}_M$ at the 1%, 5%, and 10% levels are 0.005, 0.053, and 0.107, respectively. This pattern supports the view that, under a positive measure contact set, $\hat{T}_M$ approaches the correct null size rather than exhibiting spurious over-rejection; see also Fang (2019).

Table 1: Simulation Results for Size: Rejection Rates at the 5% Level

	N1 ($\rho = 0$)				N2 ($\rho = 0.25$)			
	$n = 200$	$n = 500$	$n = 1000$	$n = 2000$	$n = 200$	$n = 500$	$n = 1000$	$n = 2000$
$\hat{T}_I$	0.044	0.043	0.041	0.059	0.056	0.040	0.054	0.043
$\hat{T}_M$	0.002	0.002	0.002	0.002	0.002	0.002	0.004	0.002
$\hat{T}_{\text{Max}}$	0.002	0.007	0.013	0.039	0.004	0.004	0.012	0.027
$\hat{T}_{\text{Sum}}$	0.006	0.004	0.005	0.008	0.008	0.013	0.010	0.013
$\hat{T}_{\text{BC}}$	0.024	0.020	0.019	0.032	0.026	0.025	0.033	0.022
	N3 ($\rho = 0.81$)				N4 (positive contact set)			
	$n = 200$	$n = 500$	$n = 1000$	$n = 2000$	$n = 200$	$n = 500$	$n = 1000$	$n = 2000$
$\hat{T}_I$	0.046	0.050	0.058	0.046	0.039	0.035	0.049	0.056
$\hat{T}_M$	0.003	0.005	0.016	0.016	0.011	0.023	0.032	0.050
$\hat{T}_{\text{Max}}$	0.003	0.005	0.019	0.021	0.011	0.023	0.032	0.048
$\hat{T}_{\text{Sum}}$	0.008	0.011	0.026	0.017	0.015	0.021	0.034	0.058
$\hat{T}_{\text{BC}}$	0.018	0.026	0.040	0.027	0.023	0.023	0.047	0.057

Power Performance for Index Sufficiency

To examine power against violations of the index sufficiency restriction, we consider four alternatives, denoted ALT1(I)–ALT4(I), that are designed to primarily violate index sufficiency while leaving the nesting inequalities approximately satisfied on the residual grid used in the simulations. Following the simulation logic in Carr and Kitagawa (2021), these designs are constructed by modifying the treated potential outcome equation within the Roy selection framework so that the distribution of $Y(1)$ depends on the instrument Z_1 beyond its effect through the propensity score P.

$$\begin{aligned}
&\text{ALT1(I)}\colon Y(1) = X\beta_1 + V(1) + 1.2 \times (1 - \Phi(Z_1)),\\
&\text{ALT2(I)}\colon Y(1) = X\beta_1 + 0.2 \times \Phi(Z_1)V(1) + (1 - \Phi(Z_1))V(1),\\
&\text{ALT3(I)}\colon Y(1) = X\beta_1 + \Phi(Z_1)V(1) + 0.2 \times (1 - \Phi(Z_1))V(1),\\
&\text{ALT4(I)}\colon Y(1) = X\beta_1 + \Phi(Z_1)V(1) + (1 - \Phi(Z_1))\mu_I,
\end{aligned} \tag{4.2}$$

where $\mu_I = A_I$, with $A_I \in \{-3, -1.5, 0, 1.5, 3\}$ drawn with probabilities $(0.15, 0.2, 0.3, 0.2, 0.15)$ independently of the remaining simulation primitives entering (4.2).

In all four designs, the conditional distribution of $Y(1)$ varies with Z_1 even after conditioning on the propensity score, so index sufficiency fails by construction. At the same time, the nesting inequalities are nearly satisfied on the residual grid used in the simulations.[4] In this sense, ALT1(I)–ALT4(I) play the same role as the index sufficiency alternatives emphasized by Carr and Kitagawa

[4]Numerical checks on the simulation residual grid show that the maximum violation of the nesting inequalities is below 0.02 across ALT1(I)–ALT4(I). These designs should therefore be interpreted as alternatives that primarily target index sufficiency, while leaving the monotonicity component close to its null.

(2021), but are formulated here directly in a continuous instrument Roy environment.

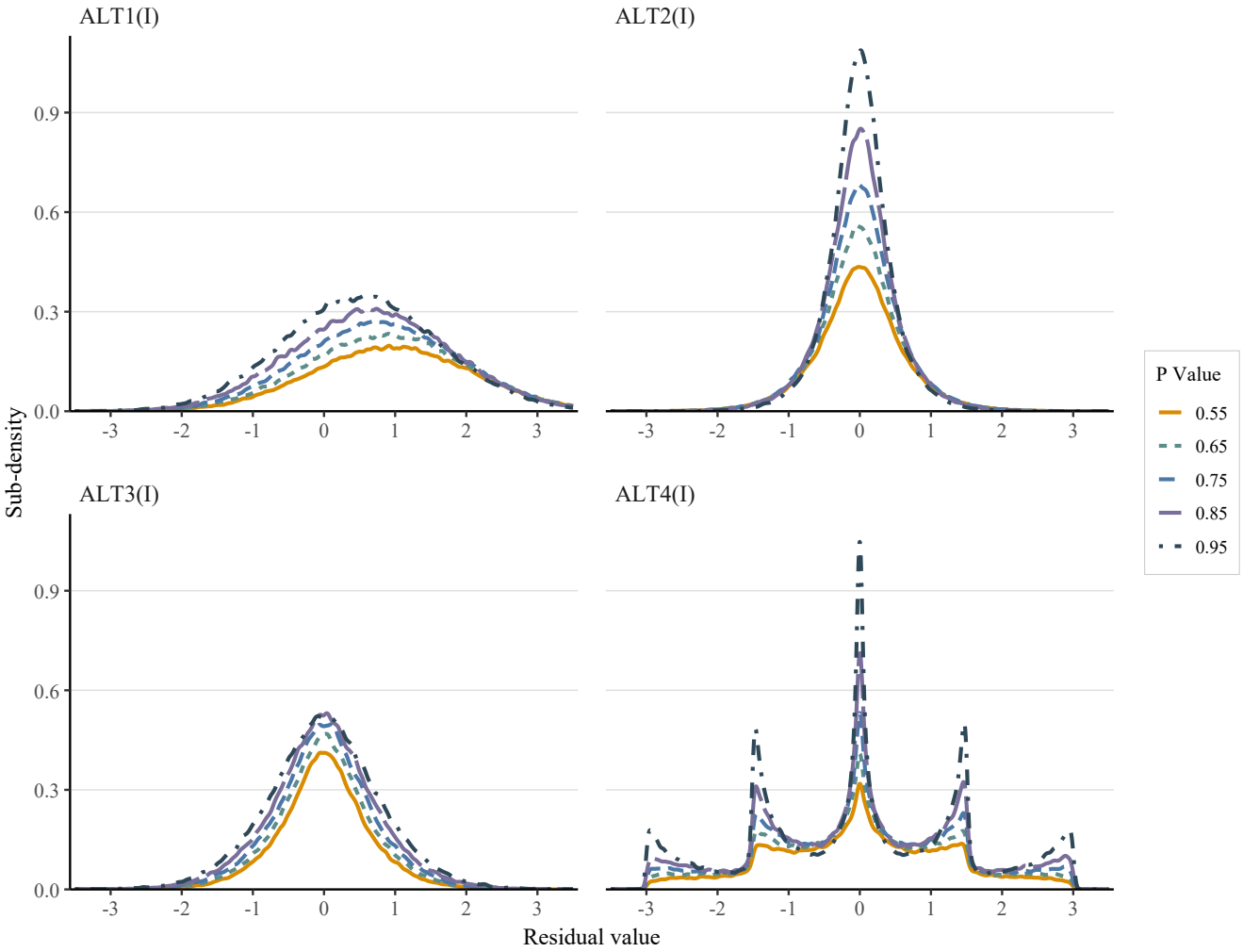
Table 2: Power under Violations of H_0^I: Rejection Rates at the 5% Level

	ALT1(I)				ALT2(I)			
	$n = 200$	$n = 500$	$n = 1000$	$n = 2000$	$n = 200$	$n = 500$	$n = 1000$	$n = 2000$
$\hat{T}_I$	0.178	0.535	0.876	0.998	0.106	0.227	0.646	0.991
$\hat{T}_M$	0.005	0.000	0.006	0.025	0.002	0.005	0.006	0.010
$\hat{T}_{\text{Max}}$	0.020	0.205	0.740	0.998	0.011	0.044	0.420	0.982
$\hat{T}_{\text{Sum}}$	0.041	0.212	0.613	0.971	0.022	0.049	0.232	0.843
$\hat{T}_{\text{BC}}$	0.125	0.428	0.807	0.998	0.060	0.135	0.498	0.974
	ALT3(I)				ALT4(I)			
	$n = 200$	$n = 500$	$n = 1000$	$n = 2000$	$n = 200$	$n = 500$	$n = 1000$	$n = 2000$
$\hat{T}_I$	0.062	0.102	0.313	0.802	0.047	0.051	0.122	0.457
$\hat{T}_M$	0.000	0.005	0.005	0.012	0.002	0.004	0.003	0.000
$\hat{T}_{\text{Max}}$	0.002	0.015	0.107	0.690	0.003	0.010	0.066	0.390
$\hat{T}_{\text{Sum}}$	0.014	0.026	0.107	0.467	0.005	0.007	0.032	0.136
$\hat{T}_{\text{BC}}$	0.038	0.055	0.186	0.692	0.029	0.025	0.081	0.319

Notes: The table reports rejection rates at the 5% level. For the multimodal designs ALT4(I) and ALT4(M), we use the auxiliary bandwidth $\ell_n = 0.5n^{-1/3}$. This design-specific choice is included for comparability with the multimodal monotonicity design and has negligible effect on the qualitative power conclusions under ALT4(I). It is not required for implementing the procedure more generally.

Table 2 reports rejection rates at the 5% level. The index statistic $\hat{T}_I$ displays clear power against all four alternatives, with rejection rates increasing monotonically in n. The gains are especially strong under ALT1(I) and ALT2(I), where the rejection rates rise from 0.178 and 0.106 at $n = 200$ to 0.998 and 0.991 at $n = 2000$, respectively. Power also increases under ALT3(I) and ALT4(I), although more gradually, reaching 0.802 and 0.457 at $n = 2000$. By contrast, the monotonicity statistic $\hat{T}_M$ shows negligible rejection rates throughout, ranging from 0.000 to 0.025 across all designs and sample sizes. This is consistent with the small numerical departures from the nesting inequalities documented above, which leave the monotonicity component close to its null value in these designs.

Figure 1: Sub-densities of (V, D) at $P = 0.55, 0.65, 0.75, 0.85$, and 0.95 under ALT1–ALT4(I)



Notes: This figure plots numerically estimated sub-densities for the four alternative simulation designs ALT1–ALT4(I), using a sample of size $n = 100{,}000$ and bandwidth $h = 0.03$. Visual violations of the nesting inequalities are not apparent on this grid, and numerical checks show only very small departures. These designs should therefore be interpreted as alternatives that primarily target index sufficiency rather than as designs that analytically impose the monotonicity null.

The three combined procedures, $\hat{T}_{\text{Sum}}$, $\hat{T}_{\text{Max}}$, and $\hat{T}_{\text{BC}}$, also gain power against these alternatives, but their finite-sample performance differs substantially across statistics. Especially under ALT1(I) and ALT2(I), their rejection rates approach one as the sample size grows, reflecting the increasing informativeness of the index component in large samples. At the same time, the differences across the combined procedures reveal a tension between the two components. In small samples, $\hat{T}_M$ still has a relatively large influence on $\hat{T}_{\text{Max}}$ and $\hat{T}_{\text{Sum}}$ under these index sufficiency alternatives, thereby diluting the power gains from $\hat{T}_I$. As a result, these two joint statistics are less powerful than the Bonferroni procedure in small and moderate samples. As n increases, $\hat{T}_I$ becomes more informative, allowing

$\hat{T}_{\text{Max}}$ to catch up to $\hat{T}_{\text{BC}}$ under the stronger alternatives, whereas $\hat{T}_{\text{Sum}}$ remains comparatively less powerful.

Power Performance for Monotonicity

To examine power against violations of the monotonicity restriction, we consider four alternatives, denoted ALT1(M)–ALT4(M), that maintain index sufficiency. Following the spirit of K15, these designs are constructed to generate distinct patterns of violations of the nesting inequalities within the same Roy selection framework.

$$\begin{aligned}
&\text{ALT1(M): } Y(1) = X\beta_1 + V(1) + \frac{1}{1.25 - P^2}, \\
&\text{ALT2(M): } Y(1) = X\beta_1 + \frac{0.25}{P^4 + 0.1}V(1), \\
&\text{ALT3(M): } Y(1) = X\beta_1 + (P^2 + 0.1)V(1), \\
&\text{ALT4(M): } Y(1) = X\beta_1 + \mu_M + 0.125V(1),
\end{aligned} \tag{4.3}$$

Here, $\mu_M = A_M P^2$, with $A_M \in \{-2, -1, 0, 1, 2\}$ drawn with probabilities $(0.15, 0.2, 0.3, 0.2, 0.15)$ independently of the remaining simulation primitives entering (4.3). By construction, these specifications preserve index sufficiency, since the treated potential outcome depends on the instruments only through the propensity score. Their purpose is instead to induce failures of the nesting inequalities of the kind emphasized by K15, but now within a continuous index MTE design. As illustrated in Figure 2, ALT1(M)–ALT4(M) generate distinct patterns of monotonicity violation.

More specifically, ALT1(M) introduces a location shift, so the violation is concentrated in the left tail. ALT2(M) changes tail thickness through a nonlinear scale effect, producing violations in the tails. ALT3(M) increases the mass at the center of the distribution, generating a more localized departure from monotonicity. ALT4(M) introduces a multimodal component and yields a broader violation pattern over the support. To avoid excessively heavy-tailed realizations induced by the nonlinear terms, we truncate the disturbance term $Y(1) - X\beta_1$ to the interval $[-3, 3]$ in ALT1(M) and ALT2(M).

Table 3: Power under Violations of H_0^M: Rejection Rates at the 5% Level

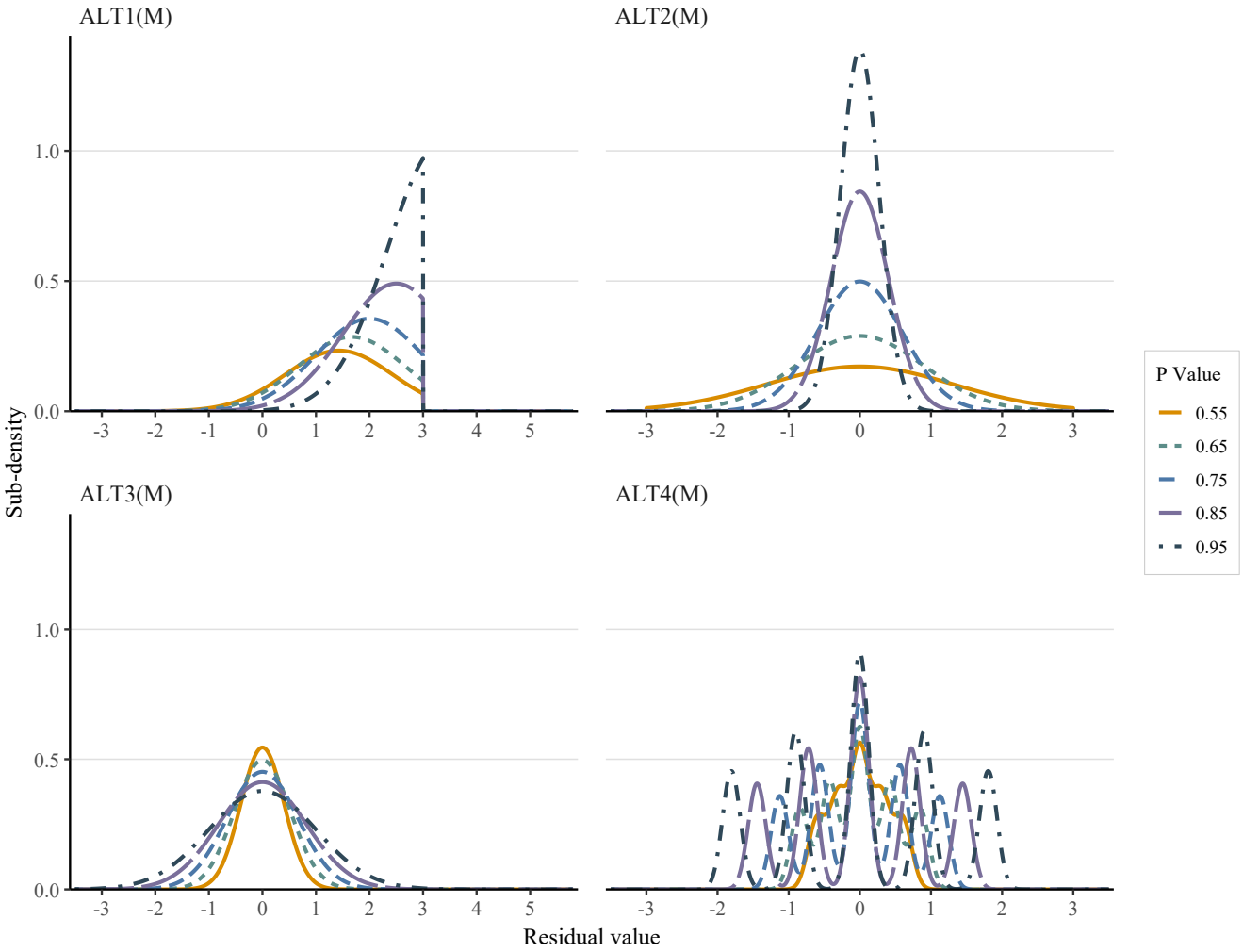
	ALT1(M)				ALT2(M)			
	$n = 200$	$n = 500$	$n = 1000$	$n = 2000$	$n = 200$	$n = 500$	$n = 1000$	$n = 2000$
$\hat{T}_I$	0.062	0.053	0.041	0.057	0.060	0.056	0.046	0.054
$\hat{T}_M$	0.666	0.998	1.000	1.000	0.016	0.223	0.926	1.000
$\hat{T}_{\text{Max}}$	0.665	0.998	1.000	1.000	0.017	0.201	0.883	1.000
$\hat{T}_{\text{Sum}}$	0.532	0.991	1.000	1.000	0.029	0.155	0.687	0.998
$\hat{T}_{\text{BC}}$	0.528	0.998	1.000	1.000	0.041	0.152	0.818	1.000
	ALT3(M)				ALT4(M)			
	$n = 200$	$n = 500$	$n = 1000$	$n = 2000$	$n = 200$	$n = 500$	$n = 1000$	$n = 2000$
$\hat{T}_I$	0.060	0.055	0.042	0.067	0.043	0.042	0.056	0.055
$\hat{T}_M$	0.010	0.084	0.207	0.491	0.010	0.027	0.100	0.460
$\hat{T}_{\text{Max}}$	0.010	0.081	0.211	0.474	0.011	0.029	0.096	0.390
$\hat{T}_{\text{Sum}}$	0.024	0.056	0.157	0.368	0.013	0.025	0.076	0.232
$\hat{T}_{\text{BC}}$	0.027	0.071	0.151	0.384	0.028	0.034	0.067	0.318

Notes: The table reports rejection rates at the 5% level. For the multimodal designs ALT4(I) and ALT4(M), we use the auxiliary bandwidth $\ell_n = 0.5n^{-1/3}$. The main motivation for this design-specific choice is to stabilize the finite-sample behavior of the index statistic $\hat{T}_I$ under ALT4(M), where index sufficiency holds. This adjustment is not needed for implementing the procedure more generally and does not drive the qualitative power patterns of the monotonicity statistics.

Table 3 reports rejection rates at the 5% level. Across all four alternatives, the index statistic $\hat{T}_I$ remains close to nominal size, with rejection rates ranging from 0.042 to 0.062. This is consistent with the design of ALT1(M)–ALT4(M), under which index sufficiency is maintained and only the monotonicity restriction fails.

The monotonicity statistic $\hat{T}_M$ shows substantial power, but the rate at which power rises varies markedly across designs. Under ALT1(M), where the violation is strong and global, rejection rates increase from 0.666 at $n = 200$ to essentially one by $n = 1000$. Under ALT2(M), power also rises sharply, from 0.016 at $n = 200$ to 0.223 at $n = 500$, 0.926 at $n = 1000$, and 1.000 at $n = 2000$. By contrast, power accumulates more gradually under ALT3(M) and ALT4(M), reaching 0.491 and 0.460, respectively, at $n = 2000$. These patterns are consistent with the more localized and diffuse nature of the monotonicity violations in those designs.

Figure 2: Sub-densities of (V, D) at $P = 0.55, 0.65, 0.75, 0.85$, and 0.95 under ALT1–ALT4(M)



Notes: This figure plots the analytically derived sub-densities for the four alternative simulation designs ALT1–ALT4(M). Violations of the nesting inequalities arise under these designs, although the form of the violation differs across alternatives.

The combined procedures also display power against monotonicity violations. Their rejection rates are largely driven by $\hat{T}_M$ in these designs, while the index component remains close to nominal size. This pattern is most visible under ALT1(M) and ALT2(M). Under ALT1(M), all three combined procedures approach power one by $n = 500$ or $n = 1000$. Under ALT2(M), $\hat{T}_{\text{Max}}$ rises from 0.017 at $n = 200$ to 0.883 at $n = 1000$ and 1.000 at $n = 2000$, closely tracking the increase in $\hat{T}_M$. For $\hat{T}_{\text{Sum}}$ and $\hat{T}_{\text{BC}}$, power also rises to one, or nearly one, by $n = 2000$.

For the more localized violations in ALT3(M) and ALT4(M), power accumulates more gradually, but the combined procedures still move in the same direction as the monotonicity statistic. At $n = 2000$, for example, $\hat{T}_{\text{Max}}$ reaches 0.474 under ALT3(M) and 0.390 under ALT4(M), while $\hat{T}_{\text{BC}}$ reaches 0.384 and 0.318, respectively. The sum statistic is somewhat less powerful in these two designs, reaching 0.368 and 0.232 at $n = 2000$, which indicates that finite sample performance can

still depend on how the index and monotonicity components are combined. Overall, however, the main message is that all the joint procedures are able to inherit the power of the monotonicity statistic when violations are concentrated in H_0^M.

5 Empirical Application

We revisit the college attendance application in Carneiro et al. (2011), a leading empirical MTE analysis of college returns in the US. This application provides a natural setting for our test because it combines a rich set of excluded instruments and covariates with an estimated propensity score that spans nearly the full unit interval. While these features are favorable for semiparametric MTE estimation, they also make the identifying assumptions substantively nontrivial: when many instruments enter the selection equation, it is not automatic that treatment choice is summarized by a single index. Our empirical analysis therefore asks whether the IV identifying assumptions underlying this widely used MTE specification are supported by the data.

The outcome variable Y is log wage in 1991, and the treatment variable D is an indicator for college attendance. The analysis uses a sample of white males drawn from the National Longitudinal Survey of Youth 1979 (NLSY), with 1,747 observations: 865 treated units and 882 control units. The treatment group ($D = 1$) includes individuals with some college education, college graduates, and postgraduates, whereas the control group ($D = 0$) consists of high school dropouts and high school graduates.

Table 4 summarizes the outcome, treatment, regressors X, and excluded instruments Z_0. The baseline specification includes 10 regressors and 4 excluded instruments, college proximity (`pub4`), local tuition costs (`tuit4c`), and local labour market conditions (`lwage5_17`, `lurate_17`)—drawn from Carneiro et al. (2011). Following their empirical specification, the first stage propensity score model augments these variables with a selected set of interactions and quadratic terms, yielding 34 predictors $Z = (X, Z_0)$ in total, consisting of 22 nonconstant covariates and 12 excluded instruments. The exact variable construction procedure is documented in the replication files.

The propensity score is estimated by Probit, and the outcome residuals are obtained from the Robinson (1988) partially linear estimator using Nadaraya–Watson kernel regression with the common bandwidth h described in Section 3. As a descriptive check, the estimated propensity score ranges from 0.003 to 1.000 after rounding, and the central 90% of its empirical rank distribution corresponds to propensity score values between 0.062 and 0.967. This region contains 775 treated and 797 control observations, indicating substantial overlap away from the extreme tails.

We assess robustness to bandwidth choice by varying $h \in \{0.3, 0.4, \ldots, 0.8\}\, n^{-1/3}$, a range satisfying the undersmoothing condition required for $\sqrt{n}$-consistency, and vary the numerical delta bootstrap tuning parameter over $\kappa \in \{0.1, 0.15, 0.2\} \log(n)\, n^{-1/2}$. Both test statistics are evaluated on a 1% percentile grid, and bootstrap p-values are computed using $B = 1{,}000$ multiplier replications with Mammen (1993) weights. Under our baseline implementation, each bandwidth–tuning combi-

nation takes about one hour on a standard laptop. Reducing the grid to 50 percentile points lowers runtime to about 15 minutes while yielding very similar results.[5]

Table 4: Definitions of the Variables Used in the Empirical Analysis

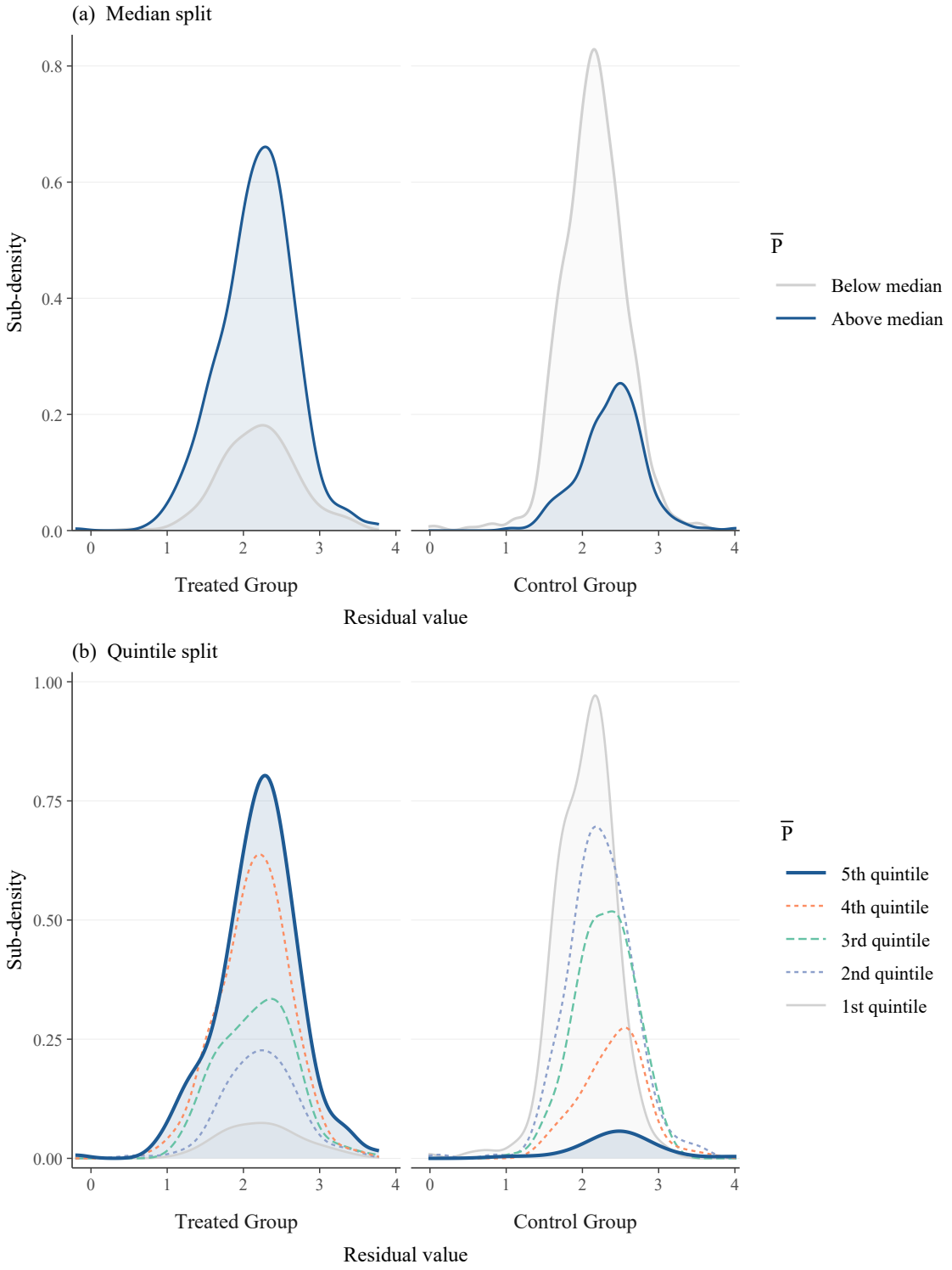
Variable	**Definition**
Y	Log wage in 1991 (average of all non-missing wages between 1989 and 1993)
D	If ever enrolled in college by 1991; zero otherwise
X	AFQT,a mother's education, number of siblings, average log earnings 1979–2000 in county of residence at 17, average unemployment 1979–2000 in state of residence at 17, urban residence at 14, cohort dummies, years of experience in 1991, average local log earnings in 1991, local unemployment in 1991
Z_0	Presence of a college at age 14 (Card 1995; Cameron and Taber 2004) Local earnings at 17 (Cameron and Heckman 1998; Cameron and Taber 2004) Local unemployment at 17 (Cameron and Heckman 1998), Local tuition in public four-year colleges at 17 (Kane and Rouse 1995)

Before turning to the formal results, we report an informal graphical check in the spirit of K15. Figure 3 plots sub-densities of residuals conditional on a coarsened propensity score $\bar{P}$ under two discretization schemes. Panel (a) uses a median split of the estimated propensity score, while Panel (b) partitions it into quintiles. Under the median split, the plots reveal no obvious violations of the nesting inequalities. Under the quintile split, by contrast, a mild crossing appears in the control-group sub-densities. The comparison across panels shows that informal assessment is sensitive to how the continuous propensity score is discretized and therefore does not yield a clear conclusion. This is precisely the setting in which a formal test that operates directly on the continuous propensity score is most useful.

To move beyond visual inspection, we conduct our formal test of the maintained semiparametric MTE specification. Table 5 reports bootstrap p-values for $\hat{T}_{\text{Max}}$, $\hat{T}_{\text{Sum}}$, $\hat{T}_I$, and $\hat{T}_M$ across all bandwidth–tuning combinations described above. Varying h affects both $\hat{T}_I$ and $\hat{T}_M$, whereas varying κ affects only $\hat{T}_M$.

[5]The computation is implemented in our R package, `falsifyMTE`. Replication code including the R package, simulation scripts, and the implementation routines will be provided as supplementary material with the final submission.

Figure 3: Sub-densities of (V, D) under coarsened propensity scores



Notes: This figure reports kernel estimates of the sub-densities of (V, D) conditional on a coarsened propensity score $\bar{P}$ using the data from Carneiro et al. (2011). Panel (a) constructs $\bar{P}$ by splitting the estimated propensity score P at its sample median. Panel (b) constructs $\bar{P}$ by partitioning the estimated propensity score P into quintiles. In both panels, the estimates use a Gaussian kernel with a rule-of-thumb bandwidth.

Across bandwidth and tuning choices, the results consistently provide no statistical evidence against the maintained semiparametric MTE specification in Carneiro et al. (2011). The component tests for index sufficiency $\hat{T}_I$ and monotonicity $\hat{T}_M$ remain well above conventional significance levels. Three finite sample patterns are worth noting. First, the p-values for $\hat{T}_M$ tend to increase with the numerical delta tuning parameter κ. Second, both component tests tend to deliver somewhat smaller p-values for intermediate bandwidths, especially around $h = 0.5n^{-1/3}$, than for smaller or larger bandwidth choices. Third, the joint statistics provide little additional information beyond the marginal component tests in this application. In every one of the 18 bandwidth–tuning combinations, the bootstrap p-value for $\hat{T}_{\text{Max}}$ is numerically identical to that for $\hat{T}_M$, and the p-value for $\hat{T}_{\text{Sum}}$ likewise tracks $\hat{T}_M$ very closely. This pattern indicates that the substantive information in Table 5 should be read primarily from the component statistics rather than from the unweighted joint statistics. This scale sensitivity is consistent with the discussion in Remark 3.1 and suggests that future work should consider studentization or alternative weighting schemes that allow the index and monotonicity components to contribute more evenly to the joint statistic.

Table 5: Empirical Application: Test Results

	$\hat{\kappa} = 0.0179$				$\hat{\kappa} = 0.0268$				$\hat{\kappa} = 0.0357$			
$\hat{h}$	$\hat{T}_{\text{Max}}$	$\hat{T}_{\text{Sum}}$	$\hat{T}_M$	$\hat{T}_I$	$\hat{T}_{\text{Max}}$	$\hat{T}_{\text{Sum}}$	$\hat{T}_M$	$\hat{T}_I$	$\hat{T}_{\text{Max}}$	$\hat{T}_{\text{Sum}}$	$\hat{T}_M$	$\hat{T}_I$
1	0.285	0.284	0.285	0.385	0.335	0.334	0.335	0.385	0.512	0.505	0.512	0.385
2	0.178	0.167	0.178	0.330	0.190	0.173	0.190	0.330	0.313	0.293	0.313	0.330
3	0.184	0.164	0.184	0.279	0.200	0.183	0.200	0.279	0.316	0.287	0.316	0.279
4	0.199	0.164	0.199	0.252	0.211	0.183	0.211	0.252	0.331	0.289	0.331	0.252
5	0.188	0.165	0.188	0.255	0.195	0.184	0.195	0.255	0.333	0.291	0.333	0.255
6	0.266	0.244	0.266	0.278	0.287	0.274	0.287	0.278	0.459	0.419	0.459	0.278

Notes: All p-values are obtained from 1,000 bootstrap iterations. The bandwidth parameter $\hat{h} \in \{1, \ldots, 6\}$ corresponds to $\{0.3n^{-1/3}, \ldots, 0.8n^{-1/3}\}$. The tuning parameter $\hat{\kappa} \in \{0.0179, 0.0268, 0.0357\}$ corresponds to $\{0.1\log(n)n^{-1/2}, 0.15\log(n)n^{-1/2}, 0.2\log(n)n^{-1/2}\}$.

6 Conclusion

This paper develops a unified specification test for instrumental variable validity in semiparametric MTE models. We show that index sufficiency and stochastic monotonicity jointly provide a *sharp* characterization of the IV identifying assumptions under the maintained semiparametric specification, including the linear outcome structure and the correctly specified propensity score. In this sense, these restrictions exhaust the information in the observed data distribution that is useful for refuting the maintained model. Building on these implications, we propose an asymptotically valid testing procedure based on a multiplier bootstrap that accommodates the generated regressor nature of both

the propensity score and the outcome residuals.

Monte Carlo simulations demonstrate that the procedure delivers accurate size control and good power against a range of violations of both index sufficiency and monotonicity. In our empirical application to the college attendance setting of Carneiro et al. (2011), we find no statistical evidence against either component of the testable implications. To our knowledge, this constitutes the first formal joint assessment of all the identifying assumptions in this widely used MTE specification.

Several directions for future work emerge naturally. First, the optimal weighting of the index and monotonicity components in joint testing merits further investigation, especially given the non-standard limit behavior of $\hat{T}_M$ when the contact set has measure zero. Second, our approach naturally complements recent work on testing IV validity in judge leniency designs (Coulibaly et al., 2024); adapting our joint testing strategy to such settings is a promising direction.

References

Beare, Brendan K. and Xiaoxia Shi, "An improved bootstrap test of density ratio ordering," *Econometrics and Statistics*, 8 2018.

Carneiro, Pedro, James J Heckman, and Edward J Vytlacil, "Estimating Marginal Returns to Education," *American Economic Review*, 10 2011, *101* (6), 2754–2781.

Carr, Thomas and Toru Kitagawa, "Testing Instrument Validity with Covariates," *Working Paper*, 12 2021.

Coulibaly, Mohamed, Yu-Chin Hsu, Ismael Mourifié, and Yuanyuan Wan, "A Sharp Test for the Judge Leniency Design," NBER Working Paper 32456, National Bureau of Economic Research 2024.

Delgado, Miguel A. and Juan Carlos Escanciano, "Distribution-free tests of stochastic monotonicity," *Journal of Econometrics*, 9 2012, *170* (1), 68–75.

_ and _ , "Conditional Stochastic Dominance Testing," *Journal of Business & Economic Statistics*, 2013, *31* (1).

_ and Wenceslao González González-Manteiga, "Significance testing in nonparametric regression based on the bootstrap," *Annals of Statistics*, 10 2001, *29* (5), 1469–1507.

Dobbie, Will, Jacob Goldin, and Crystal S. Yang, "The Effects of Pretrial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges†," *American Economic Review*, 2018, *108* (2), 201–240.

Escanciano, Juan Carlos, David T. Jacho-Chávez, and Arthur Lewbel, "Uniform convergence of weighted sums of non and semiparametric residuals for estimation and testing," *Journal of Econometrics*, 2014, *178* (PART 3), 426–443.

Fang, "Refinements of the Kiefer-Wolfowitz theorem and a test of concavity," *Electronic Journal of Statistics*, 2019.

Fang, Zheng and Andres Santos, "Inference on Directionally Differentiable Functions," *The Review of Economic Studies*, 9 2018, *86* (1), 377–412.

Gupta, Atul, Sabrina T Howell, Constantine Yannelis, and Abhinav Gupta, "Owner incentives and performance in healthcare: private equity investment in nursing homes," *The Review of Financial Studies*, 2024, *37* (4), 1029–1077.

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

Heckman, James J. and Edward Vytlacil, "Structural equations, treatment effects, and econometric policy evaluation," *Econometrica*, 5 2005, *73* (3), 669–738.

Hong, Han and Jessie Li, "The numerical delta method," *Journal of Econometrics*, 2018, *206* (2), 379–394.

__ **and** __ , "The numerical bootstrap," *Annals of Statistics*, 2020, *48* (1), 397–412.

Hsu, Yu-Chin, Chu-An Liu, and Xiaoxia Shi, "Testing generalized regression monotonicity," *Econometric Theory*, 2019, *35* (6), 1146–1200.

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

Kitagawa, Toru, "A Test for Instrument Validity," *Econometrica*, 2015, *83* (5), 2043–2063.

Maistre, Samuel and Valentin Patilea, "Nonparametric model checks of single-index assumptions," *Statistica Sinica*, 2019, *29* (1), 113–138.

Mammen, Enno, "Bootstrap and wild bootstrap for high-dimensional linear models," 1993.

Mogstad, Magne and Alexander Torgovitsky, "Instrumental Variables with Unobserved Heterogeneity in Treatment Effects," in Christian Dustmann and Thomas Lemieux, eds., *Handbook of Labor Economics*, Vol. 5, Elsevier, 2024, pp. 1–114.

Mourifié, Ismael and Yuanyuan Wan, "Testing local average treatment effect assumptions," *Review of Economics and Statistics*, 2017, *99* (2), 305–313.

Mourifié, Ismael and Yuanyuan Wan, "Layered policy analysis in program evaluation using the marginal treatment effect," *Journal of Econometrics*, 2025, *251*, 106060.

Nelsen, RB, *An introduction to copulas*, Springer, 2007.

Robinson, P. M., "Root-N-Consistent Semiparametric Regression," *Econometrica*, 7 1988, *56* (4), 931.

Sant'Anna, Pedro H.C. and Xiaojun Song, "Specification tests for the propensity score," *Journal of Econometrics*, 2019, *210* (2), 379–404.

Schoenberg, Uta, Thomas Cornelissen, Christian Dustmann, and Anna Raute, "Who Benefits from Universal Child Care? Estimating Marginal Returns to Early Child Care Attendance," *Journal of Political Economy*, 8 2018, pp. 699–979.

Seo, Juwon, "Tests of stochastic monotonicity with improved power," *Journal of Econometrics*, 11 2018, *207* (1), 53–70.

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

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