北京大学国家发展研究院
National School of Development at Peking University
中国经济研究中心
China Centre for Economic Research
讨论稿系列
Working Paper Series

E2023004 2023-02-20

Instrument Validity for Heterogeneous Causal Effects

Zhenting Sun
China Center for Economic Research
National School of Development
Peking University

Abstract

This paper provides a general framework for testing instrument validity in heterogeneous causal effect models. The generalization includes the cases where the treatment can be multivalued ordered or unordered. Based on a series of testable implications, we propose a nonparametric test which is proved to be asymptotically size controlled and consistent. Compared to the tests in the literature, our test can be applied in more general settings and may achieve power improvement. Refutation of instrument validity by the test helps detect invalid instruments that may yield implausible results on causal effects. Evidence that the test performs well on finite samples is provided via simulations. We revisit the empirical study on return to schooling to demonstrate application of the proposed test in practice. An extended continuous mapping theorem and an extended delta method, which may be of independent interest, are provided to establish the asymptotic distribution of the test statistic under null.

Keywords: Instrument validity, heterogeneous causal effects, general nonparametric test, power improvement, extended continuous mapping theorem, extended delta method

Instrument Validity for Heterogeneous Causal Effects

Zhenting Sun*
China Center for Economic Research
National School of Development
Peking University
zhentingsun@nsd.pku.edu.cn

October 3, 2022

Abstract

This paper provides a general framework for testing instrument validity in heterogeneous causal effect models. The generalization includes the cases where the treatment can be multivalued ordered or unordered. Based on a series of testable implications, we propose a nonparametric test which is proved to be asymptotically size controlled and consistent. Compared to the tests in the literature, our test can be applied in more general settings and may achieve power improvement. Refutation of instrument validity by the test helps detect invalid instruments that may yield implausible results on causal effects. Evidence that the test performs well on finite samples is provided via simulations. We revisit the empirical study on return to schooling to demonstrate application of the proposed test in practice. An extended continuous mapping theorem and an extended delta method, which may be of independent interest, are provided to establish the asymptotic distribution of the test statistic under null.

Keywords: Instrument validity, heterogeneous causal effects, general nonparametric test, power improvement, extended continuous mapping theorem, extended delta method

*This article is a revised version of the first chapter of the author's doctoral thesis at UC San Diego. I am deeply grateful to Brendan K. Beare, Zheng Fang, Andres Santos, Yixiao Sun, and Kaspar Wüthrich for their constant support on this paper. I thank the editors and three anonymous referees for their constructive suggestions that help improve the paper significantly. I thank Shengtao Dai, Tongyu Li, and Xingyu Li for their excellent work as research assistants. I also thank Roy Allen, Qihui Chen, Asad Dossani, Graham Elliott, Ivan Fernandez-Val, Wenzheng Gao, James D. Hamilton, Jungbin Hwang, Toru Kitagawa, Sungwon Lee, Juwon Seo, Xiaojun Song, and all seminar participants for their insightful suggestions and comments.

1 Introduction

The local average treatment effect (LATE) framework, introduced by the seminal works of Imbens and Angrist (1994) and Angrist et al. (1996), is a commonly used approach in studies of instrumental variable (IV) models with treatment effect heterogeneity. The local quantile treatment effect (LQTE) is a concept similar to LATE. While LATE shows the treatment effect on the mean of the outcome, LQTE is more informative in regard to the effect on the outcome distribution.[1] These causal effect models rely on several strong and sometimes controversial assumptions of IV validity: 1) The instrument should not affect the outcome directly; 2) it should be as good as random assignment; and 3) it affects the treatment in monotone fashion. Violations of these conditions can generally lead to unidentification and inconsistent estimation of treatment effects. Relevant surveys and discussion of this can be found in Angrist and Pischke (2008), Angrist and Pischke (2014), Imbens (2014), Imbens and Rubin (2015), Koenker et al. (2017), Melly and Wüthrich (2017), and Huber and Wüthrich (2018). Since the plausibility of the analyses of such models depends on IV validity, economics research has developed methods to examine these conditions based on testable implications.

Cases where the IV validity conditions may be violated can be found in empirical applications. For example, the college proximity was used as an instrument of education attainment in the study of Card (1993). If the education level is treated as a binary variable (four-year college degree), the validity of the college proximity is rejected by the test of Kitagawa (2015) when no conditioning covariates are added in the model. Mogstad et al. (2021) considered tuition and college proximity as multiple instruments for college attendance. They showed that, if a homogeneity condition does not hold for individuals, the validity of the multiple instruments will be violated. The quarter of birth instrument used in Angrist and Krueger (1991) is questionable because the exclusion restriction may not hold due to seasonal birth patterns (Bound et al., 1995; Buckles and Hungerman, 2013). The monotonicity condition of IV validity fails in the selection with two-way flows example in Lee and Salanié (2018).[2]

Kitagawa (2015) was the first paper to propose a test of IV validity in heterogeneous causal effect models with a binary treatment based on the testable implications in the literature. It was the first to show the sharpness of these testable implications. Their test, constructed using a bootstrap method, was shown to be asymptotically uniformly size controlled and consistent. Mourifié and Wan (2017) reformulated the testable implications used in Kitagawa (2015) as conditional inequalities. They then showed that these inequal-

[1] See, for example, studies of LQTE in Abadie (2002), Ananat and Michaels (2008), Cawley and Meyerhoefer (2012), Frölich and Melly (2013), and Eren and Ozbeklik (2014).

[2] See further discussion on IV validity assumptions in Section 5.1.

ities could be tested in the intersection bounds framework of Chernozhukov et al. (2013) using the Stata package provided by Chernozhukov et al. (2015). The present paper provides a general framework for testing such IV validity assumptions. The proposed test can be applied in more general settings in which the treatment variable can be multivalued ordered or unordered[3] and the outcome variable can be unbounded[4]. Also, the proposed test achieves power improvement by solving a technical issue and employing a novel bootstrap approach. Huber and Mellace (2015) derived a testable implication for a weaker LATE identifying condition, that is, that the potential outcomes are mean independent of instruments, conditional on each selection type.[5] The focus of the present paper is on full statistical independence of potential outcomes and instruments.

A modified variance-weighted Kolmogorov–Smirnov (KS) test statistic is employed in our test. As mentioned by Kitagawa (2015), variance-weighted KS statistics have been widely applied in the literature on conditional moment inequalities, such as in Andrews and Shi (2013), Armstrong (2014), Armstrong and Chan (2016), and Chetverikov (2018). More general KS statistics can be found in the stochastic dominance testing literature, such as in Abadie (2002), Barrett and Donald (2003), Horváth et al. (2006), Linton et al. (2010), Barrett et al. (2014), and Donald and Hsu (2016). To investigate the asymptotic properties of the proposed test, we introduce L^r ($r \in \mathbb{N}$) spaces with which a series of fundamental results are established, such as the compactness of particular function spaces, the Glivenko–Cantelli and the Donsker results, and so on. Based on these results, we obtain the asymptotic behavior of the test statistic.[6] The asymptotic properties of the proposed test are established accordingly.

There are two major complications in deriving and approximating the asymptotic distribution of the test statistic under null. First, the test statistic involves a nonsmooth (nondifferentiable) map of unknown parameters (underlying probability distributions), and the delta method fails to work. We provide an extended continuous mapping theorem and an extended delta method, which might be of independent interest, to overcome this difficulty. By showing that the conditions of the extended delta method are satisfied under several weak assumptions, we establish the null asymptotic distribution of the test statistic. Second,

[3]Studies of LATE with binary treatments can be found in Angrist (1990), Angrist and Krueger (1991), and Vytlacil (2002). Those with multivalued treatments can be found in Angrist and Imbens (1995), Angrist and Krueger (1995), and Vytlacil (2006). Identification of causal effects in unordered choice (treatment) models can be found in Heckman et al. (2006), Heckman and Vytlacil (2007), Heckman et al. (2008), and Heckman and Pinto (2018).

[4]See Reed (2001, 2003) and Toda (2012) for the approximation of income distributions by members of the double Pareto parametric family.

[5]The condition of potential outcomes being mean independent of instruments is not sufficient if we are concerned with distributional features of a complier's potential outcomes, such as the quantile treatment effects for compliers; see Abadie et al. (2002) for details.

[6]See further discussion before Theorem 3.1.

since the null asymptotic distribution involves a nonlinear function, the standard bootstrap method may fail to approximate this distribution consistently. Discussion of this issue can be found in Dümbgen (1993), Andrews (2000), Hirano and Porter (2012), Hansen (2017), Fang and Santos (2018), and Hong and Li (2018). To achieve a consistent approximation, we extend the bootstrap approach proposed by Fang and Santos (2018)[7] and provide a valid bootstrap critical value. The test is found to be asymptotically size controlled and consistent. Evidence that the test performs well on finite samples is provided via simulations.

We now introduce the following notation, which will be used throughout the paper. We let $\rightsquigarrow$ denote Hoffmann–Jørgensen weak convergence in a metric space. For a set $\mathbb{D}$, denote the space of bounded functions on $\mathbb{D}$ by $\ell^\infty(\mathbb{D})$: $\ell^\infty(\mathbb{D}) = \{f : \mathbb{D} \to \mathbb{R} : \|f\|_\infty < \infty\}$, where $\|f\|_\infty = \sup_{x\in\mathbb{D}} |f(x)|$. If $\mathbb{D}$ is a topological space, let $C(\mathbb{D})$ denote the set of continuous functions on $\mathbb{D}$: $C(\mathbb{D}) = \{f : \mathbb{D} \to \mathbb{R} : f \text{ is continuous}\}$. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space on which all random elements are well defined. Let $\mathcal{B}_{\mathbb{R}^m}$ denote the Borel σ-algebra on $\mathbb{R}^m$ for all $m \in \mathbb{N}$. We use $\hat{Q}$ and $\hat{Q}^B$ to denote the empirical probability measure and the bootstrap empirical probability measure of each probability measure Q, respectively.

2 Setup and Testable Implications

2.1 Binary Treatment

To formally introduce the topic of interest, we first consider the heterogeneous causal effect model of Imbens and Angrist (1994). Let $Y \in \mathbb{R}$ be the observable outcome variable, and let $D \in \{0,1\}$ be the observable treatment variable, where $D = 1$ indicates that an individual receives treatment. Let $Z \in \{0,1\}$ be a binary instrumental variable. Let $Y_{dz} \in \mathbb{R}$ be the potential outcome variable[8] for $D = d$ and $Z = z$, where $d, z \in \{0,1\}$. Similarly, let D_z be the potential treatment variable for $Z = z$. The instrument validity assumption for binary treatment and binary IV is formalized as follows.

Assumption 2.1 *IV validity for binary D and binary Z:*

(i) Instrument Exclusion: For each $d \in \{0,1\}$, $Y_{d0} = Y_{d1}$ almost surely.

(ii) Random Assignment: The variable Z is jointly independent of $(Y_{00}, Y_{01}, Y_{10}, Y_{11}, D_0, D_1)$.

(iii) Instrument Monotonicity: The potential treatment response indicators satisfy $D_1 \geq D_0$ almost surely.

[7]Other applications of this bootstrap method can be found in Beare and Moon (2015), Beare and Fang (2017), Seo (2018), Beare and Shi (2019), and Sun and Beare (2021). A similar bootstrap approach can be found in Hong and Li (2018).

[8]See Rubin (1974) and Splawa-Neyman et al. (1990) for further discussion of the potential outcomes.

Assumption 2.1 is from Imbens and Rubin (1997), but it does not require strict instrument monotonicity. In this paper, we are not concerned with the strict monotonicity assumption, which is also known as the instrument relevance assumption.[9]

For all Borel sets B and C, we follow Kitagawa (2015) and define probability measures as follows:[10]

$$P_1(B,C) = \mathbb{P}(Y \in B, D \in C | Z = 1) \text{ and } P_0(B,C) = \mathbb{P}(Y \in B, D \in C | Z = 0).$$

Under Assumption 2.1(i), we can define a potential outcome variable Y_d such that $Y_d = Y_{d0} = Y_{d1}$ almost surely. Imbens and Rubin (1997) showed that for every Borel set B,

$$\begin{aligned} &P_1(B,\{1\}) - P_0(B,\{1\}) = \mathbb{P}(Y_1 \in B, D_1 > D_0) \\ &\text{and } P_0(B,\{0\}) - P_1(B,\{0\}) = \mathbb{P}(Y_0 \in B, D_1 > D_0). \end{aligned} \tag{1}$$

To see why (1) is true, we can write

$$\begin{aligned} P_1(B,\{1\}) - P_0(B,\{1\}) &= \mathbb{P}(Y \in B, D = 1 | Z = 1) - \mathbb{P}(Y \in B, D = 1 | Z = 0) \\ &= \mathbb{P}(Y_1 \in B, D_1 = 1) - \mathbb{P}(Y_1 \in B, D_0 = 1) = \mathbb{P}(Y_1 \in B, D_1 = 1, D_0 = 0), \end{aligned}$$

where the second equality follows from Assumptions 2.1(i) and 2.1(ii) and the third equality follows from Assumption 2.1(iii). Similar reasoning yields the second equation in (1). Since the probabilities in (1) are nonnegative, we obtain the testable implication of Assumption 2.1 in Balke and Pearl (1997), Imbens and Rubin (1997), and Heckman and Vytlacil (2005): For all $B \in \mathcal{B}_{\mathbb{R}}$,

$$P_1(B,\{1\}) - P_0(B,\{1\}) \geq 0 \text{ and } P_0(B,\{0\}) - P_1(B,\{0\}) \geq 0. \tag{2}$$

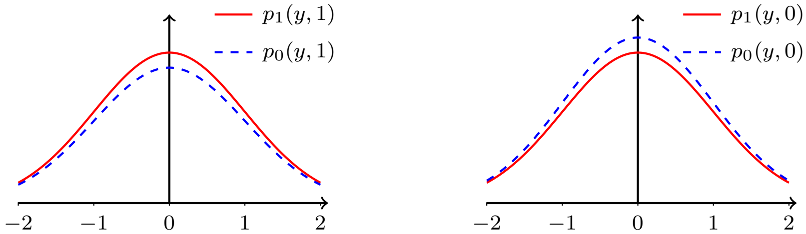
To understand (2) graphically, suppose that Y is a continuous variable and that $p_z(y,d)$ is the derivative of the function $P_z((-\infty, y], \{d\})$ with respect to y for all $d, z \in \{0,1\}$. The following graphs show a case where (2) holds.

The first inequality in (2) is shown in Figure 1a, where the derivative $p_1(y,1)$ is greater than $p_0(y,1)$ everywhere. The second inequality in (2) is shown in Figure 1b, where the derivative $p_0(y,0)$ is greater than $p_1(y,0)$ everywhere. Additional graphical examples can be found in Kitagawa (2015).

[9]As mentioned by Kitagawa (2015), the instrument relevance assumption can be assessed by inferring the coefficient in the first-stage regression of D onto Z.

[10]For simplicity of notation, we implicitly assume that (Y, D, Z) is $(\mathcal{A}, \mathcal{B}_{\mathbb{R}^3})$-measurable.

Figure 1: A special case satisfying testable implication (2)



(a) $P_1(B,\{1\}) > P_0(B,\{1\})$ (b) $P_0(B,\{0\}) > P_1(B,\{0\})$

2.2 Multivalued Ordered Treatment

Section 2.1 discussed the case where the treatment and the instrument are both binary. In many applications, D and Z can be multivalued. See, for example, Angrist and Imbens (1995), where the treatment variable is the number of years of schooling completed by a student and can take more than two values. Now suppose that $D \in \mathcal{D} = \{d_1, \ldots, d_J\}$[11] and $Z \in \mathcal{Z} = \{z_1, \ldots, z_K\}$. We let d_{max} be the maximum value of D, and d_{min} the minimum value of D. Suppose there exist potential variables Y_{dz} for $d \in \mathcal{D}$ and $z \in \mathcal{Z}$, and D_z for $z \in \mathcal{Z}$. The IV validity assumption for multivalued treatment D and multivalued instrument Z is then formalized as follows.

Assumption 2.2 *IV validity for multivalued D and multivalued Z:*

(i) Instrument Exclusion: For all $d \in \mathcal{D}$, $Y_{dz_1} = Y_{dz_2} = \cdots = Y_{dz_K}$ almost surely.

(ii) Random Assignment: The variable Z is jointly independent of $(\tilde{Y}, \tilde{D})$, where

$$\tilde{Y} = (Y_{d_1 z_1}, \ldots, Y_{d_1 z_K}, \ldots, Y_{d_J z_1}, \ldots, Y_{d_J z_K}) \text{ and } \tilde{D} = (D_{z_1}, \ldots, D_{z_K}).$$

(iii) Instrument Monotonicity: The potential treatment response variables satisfy $D_{z_{k+1}} \geq D_{z_k}$ almost surely for all $k \in \{1, 2, \ldots, K-1\}$.

Assumption 2.2 is similar to Assumptions 1 and 2 of Angrist and Imbens (1995). Theorems 1 and 2 of Angrist and Imbens (1995) showed that a weighted average of K average causal responses can be identified under Assumption 2.2. Since we allow multivalued Z, the monotonicity assumption needs to hold for each pair $(D_{z_k}, D_{z_{k+1}})$. The next lemma establishes a testable implication of Assumption 2.2.

[11] The support $\mathcal{D}$ can be generalized to the case where $\mathcal{D} = \{d_1, d_2, \ldots\}$. See details in the supplementary appendix.

Lemma 2.1 *A testable implication of Assumption 2.2 is that for all k with $1 \leq k \leq K-1$, all Borel sets B, and all $C = (-\infty, c]$ with $c \in \mathbb{R}$, the following hold:*

$$\mathbb{P}\left(Y \in B, D = d_{\max} | Z = z_k\right) \leq \mathbb{P}\left(Y \in B, D = d_{\max} | Z = z_{k+1}\right)$$
$$\text{and } \mathbb{P}\left(Y \in B, D = d_{\min} | Z = z_k\right) \geq \mathbb{P}\left(Y \in B, D = d_{\min} | Z = z_{k+1}\right); \quad (3)$$
$$\mathbb{P}\left(D \in C | Z = z_k\right) \geq \mathbb{P}\left(D \in C | Z = z_{k+1}\right). \quad (4)$$

Lemma 2.1 generalized testable implication (2) to the case where the treatment and the instrument can both be multivalued. The testable implication (first-order stochastic dominance) discussed by Angrist and Imbens (1995) for Assumption 2.2 is equivalent to (4). Clearly, if D and Z are both binary as assumed in Section 2.1, with $d_{\max} = 1$ and $d_{\min} = 0$, then (3) is equivalent to (2) and (4) is implied by (3). Liu et al. (2020) proposed testable implication (3) for the case where $\mathcal{D} = \{0, 1, 2\}$ and $\mathcal{Z} = \{0, 1, 2\}$, and (4) is also implied by (3) in this case. Thus, (3) and (4) together can be viewed as a generalized form of their condition.

2.3 Unordered Treatment

Studies of identification of causal effects in unordered choice (treatment) models can be found in Heckman et al. (2006), Heckman and Vytlacil (2007), and Heckman et al. (2008). Heckman and Pinto (2018) showed that the assumptions[12] in the preceding literature could be relaxed, and they defined a new monotonicity condition for the identification of causal effects in such models. We follow Heckman and Pinto (2018) and suppose that the support $\mathcal{D}$ of D is an unordered set with $\mathcal{D} = \{d_1, \ldots, d_J\}$ and that the support $\mathcal{Z}$ of Z with $\mathcal{Z} = \{z_1, \ldots, z_K\}$ can be unordered as well. The unordered monotonicity condition proposed by Heckman and Pinto (2018) is as follows (Assumption A-3 of Heckman and Pinto (2018)).

Assumption 2.3 *The potential treatment response indicators satisfy the condition that for all $d \in \mathcal{D}$ and all $z, z' \in \mathcal{Z}$, $1\{D_{z'} = d\} \geq 1\{D_z = d\}$ almost surely or $1\{D_{z'} = d\} \leq 1\{D_z = d\}$ almost surely.*

It is worth noting that in Assumption 2.3, D is allowed to be a vector random element. In the case where $D, Z \in \{0, 1\}$, Assumption 2.3 is equivalent to the assumption that $1\{D_1 = 1\} \geq 1\{D_0 = 1\}$ almost surely or $1\{D_1 = 1\} \leq 1\{D_0 = 1\}$ almost surely. In practice, we often assume a specific direction in the assumption, such as $1\{D_1 = 1\} \geq 1\{D_0 = 1\}$ almost surely, which is equivalent to $D_1 \geq D_0$ almost surely in Assumption 2.1(iii). With the specific direction, we can prespecify a set $\mathcal{C} \subset \mathcal{D} \times \mathcal{Z} \times \mathcal{Z}$ and assume that $1\{D_{z'} = d\} \leq 1\{D_z = d\}$ almost surely for all $(d, z, z') \in \mathcal{C}$. For example, in the above case where $D, Z \in$

[12] See Heckman and Pinto (2018, pp. 2–3) for a discussion of these assumptions.

$\{0,1\}$ and $1\{D_1 = 1\} \geq 1\{D_0 = 1\}$ almost surely, we let $\mathcal{C} = \{(0,0,1),(1,1,0)\}$. With this monotonicity condition of specified direction, we introduce the IV validity assumption for unordered treatment.[13]

Assumption 2.4 *IV validity for unordered D and unordered Z:*

(i) Instrument Exclusion: For all $d \in \mathcal{D}$ and all $z, z' \in \mathcal{Z}$, $Y_{dz} = Y_{dz'}$ almost surely.

(ii) Random Assignment: The random element Z is jointly independent of $(\tilde{Y}, \tilde{D})$, where

$$\tilde{Y} = (Y_{d_1 z_1}, \ldots, Y_{d_1 z_K}, \ldots, Y_{d_J z_1}, \ldots, Y_{d_J z_K}) \text{ and } \tilde{D} = (D_{z_1}, \ldots, D_{z_K}).$$

(iii) Instrument Monotonicity: The potential treatment elements satisfy the condition that $1\{D_{z'} = d\} \leq 1\{D_z = d\}$ almost surely for all $(d, z, z') \in \mathcal{C}$.

Under this assumption, we can define Y_d such that $Y_d = Y_{dz}$ almost surely for all z, and hence

$$\begin{aligned}\mathbb{P}\left(Y \in B, D = d | Z = z'\right) &= E[1\{Y_d \in B\} \cdot 1\{D_{z'} = d\}] \\ &\leq E[1\{Y_d \in B\} \cdot 1\{D_z = d\}] = \mathbb{P}\left(Y \in B, D = d | Z = z\right)\end{aligned}$$

for all Borel sets B and all $(d, z, z') \in \mathcal{C}$.

Lemma 2.2 *A testable implication of Assumption 2.4 is given by*

$$\mathbb{P}\left(Y \in B, D = d | Z = z'\right) \leq \mathbb{P}\left(Y \in B, D = d | Z = z\right) \tag{5}$$

for all Borel sets B and all $(d, z, z') \in \mathcal{C}$, where $\mathcal{C}$ is a prespecified subset of $\mathcal{D} \times \mathcal{Z} \times \mathcal{Z}$.

As shown in Kitagawa (2015) and Mourifié and Wan (2017), the testable implication in (2) is sharp. When the treatment or the instrument is multivalued (ordered or unordered), the cases could be complicated. Kédagni and Mourifié (2020) considered testing the joint assumptions of instrument exclusion and statistical independence, which are parts of (and different from) Assumption 2.2 and Assumption 2.4. The exclusion condition of Kédagni and Mourifié (2020) is the same as that in the present paper (Assumption 2.2(i) and Assumption 2.4(i)). The statistical independence condition of Kédagni and Mourifié (2020) (the instrument Z is jointly independent of $(Y_{d_1}, \ldots, Y_{d_J})$) is weaker than (and implied by)

[13] The test proposed in this paper can be extended for Assumption 2.3 in which the direction is not specified. See details in Appendix D.

the random assignment condition in the present paper (Assumption 2.2(ii) and Assumption 2.4(ii)). Thus, the underlying assumptions tested by Kédagni and Mourifié (2020) are weaker than those tested by the present paper.

Kédagni and Mourifié (2020) provided sharp testable implications (the generalized instrumental inequalities) for the joint assumptions of instrument exclusion and statistical independence. Consider a simple case where the outcome $Y \in \{0,1\}$, the treatment $D \in \mathcal{D}$ is multivalued, and the instrument $Z \in \mathcal{Z}$ is also multivalued. Suppose that the exclusion condition and the statistical independence condition hold. We can then define $Y_d = Y_{dz_1} = \cdots = Y_{dz_K}$ for every $d \in \mathcal{D}$. For each $y \in \{0,1\}$, every $d \in \mathcal{D}$, and every $z \in \mathcal{Z}$, we have that

$$\mathbb{P}(Y = y, D = d|Z = z) \leq \mathbb{P}(Y_d = y), \tag{6}$$

which implies that

$$\max_{d\in\mathcal{D}} \sum_{y\in\{0,1\}} \max_{z\in\mathcal{Z}} \mathbb{P}(Y = y, D = d|Z = z) \leq \sum_{y\in\{0,1\}} \mathbb{P}(Y_d = y) = 1. \tag{7}$$

For all $y_1, \ldots, y_J \in \{0,1\}$,

$$\mathbb{P}(Y_{d_1} = y_1, \ldots, Y_{d_J} = y_J) = \min_{z\in\mathcal{Z}} \mathbb{P}(Y_{d_1} = y_1, \ldots, Y_{d_J} = y_J|Z = z)$$
$$= \min_{z\in\mathcal{Z}} \sum_{j=1}^{J} \mathbb{P}(Y_{d_1} = y_1, \ldots, Y_{d_J} = y_J, D = d_j|Z = z) \leq \min_{z\in\mathcal{Z}} \sum_{j=1}^{J} \mathbb{P}(Y = y_j, D = d_j|Z = z).$$

It then follows that

$$\begin{aligned} &\sum_{y_1\in\{0,1\}} \cdots \sum_{y_J\in\{0,1\}} \min_{z\in\mathcal{Z}} \sum_{j=1}^{J} \mathbb{P}(Y = y_j, D = d_j|Z = z) \\ \geq &\sum_{y_1\in\{0,1\}} \cdots \sum_{y_J\in\{0,1\}} \mathbb{P}(Y_{d_1} = y_1, \ldots, Y_{d_J} = y_J) = 1. \end{aligned} \tag{8}$$

Next, for every j and every $y_j \in \{0,1\}$,

$$\begin{aligned} \mathbb{P}(Y_{d_j} = y_j) &= \sum_{y_1\in\{0,1\}} \cdots \sum_{y_{j-1}\in\{0,1\}} \sum_{y_{j+1}\in\{0,1\}} \cdots \sum_{y_J\in\{0,1\}} \mathbb{P}(Y_{d_1} = y_1, \ldots, Y_{d_J} = y_J) \\ &\leq \sum_{y_1\in\{0,1\}} \cdots \sum_{y_{j-1}\in\{0,1\}} \sum_{y_{j+1}\in\{0,1\}} \cdots \sum_{y_J\in\{0,1\}} \min_{z\in\mathcal{Z}} \sum_{\xi=1}^{J} \mathbb{P}(Y = y_\xi, D = d_\xi|Z = z). \end{aligned}$$

With (6), we have that

$$\max_{j\in\{1,\ldots,J\}}\max_{y_j\in\{0,1\}}\left\{\max_{z\in\mathcal{Z}}\mathbb{P}(Y=y_j,D=d_j|Z=z)-\varphi_j(y_j)\right\}\leq 0, \tag{9}$$

where

$$\varphi_j(y_j)=\sum_{y_1\in\{0,1\}}\cdots\sum_{y_{j-1}\in\{0,1\}}\sum_{y_{j+1}\in\{0,1\}}\cdots\sum_{y_J\in\{0,1\}}\min_{z\in\mathcal{Z}}\sum_{\xi=1}^{J}\mathbb{P}(Y=y_\xi,D=d_\xi|Z=z).$$

The inequalities in (7)–(9) are the testable restrictions derived by Kédagni and Mourifié (2020), which are different from the proposed testable implications in (3)–(5). Kédagni and Mourifié (2020) suggest using the approach of Chernozhukov et al. (2013) to test the restrictions in (7)–(9).[14]

Though the underlying assumptions tested by Kédagni and Mourifié (2020) are weaker than those tested by the present paper, no evidence has been found that, in general, the testable implications of Kédagni and Mourifié (2020) are weaker than (or implied by) those proposed by the present paper. Thus, to the best of our knowledge, the testable implications in Kédagni and Mourifié (2020) and those in the present paper could be complementary to each other. That is, the proposed testable restrictions may not be sharp for the IV validity Assumptions 2.2 and 2.4, and the proposed test may not be testing all possible restrictions. In practice, we suggest that users first apply the method of Chernozhukov et al. (2013) to test the restrictions in Kédagni and Mourifié (2020), and then apply the proposed method to test the joint IV validity assumptions in the present paper. In this way, the test results could be more informative about which part of the IV validity assumptions may fail.

Another interesting question is that if we combine the inequalities of Kédagni and Mourifié (2020) and those of the present paper together, are they sharp for the IV validity assumptions? To show this, we may draw on the sharpness results of Kitagawa (2015), Mourifié and Wan (2017), and Kédagni and Mourifié (2020). However, this would not be straightforward because we now allow both the treatment and the instrument to be multivalued (ordered or unordered), and the IV validity assumptions involve more conditions (random assignment and monotonicity). Since this technical complication may be beyond the main context of the present paper, we leave it for future study as an independent topic.

[14] See Section 5 of Kédagni and Mourifié (2020).

3 Test Formulation

To highlight the idea, we first introduce the test for the case where the treatment is multi-valued ordered, with support $\mathcal{D} = \{d_1, \ldots, d_J\}$. The unordered treatment case will be discussed as an extension in Section 3.3. Section B in the appendix extends the proposed test for the cases where conditioning covariates may be present. Also, we let Z be multivalued with support $\mathcal{Z} = \{z_1, \ldots, z_K\}$. The test is constructed based on the testable implication given in (3) and (4). Without loss of generality, we assume that $d_{\min} = 0$ and $d_{\max} = 1$. In practice, we can always normalize $d_{\min}$ and $d_{\max}$ to 0 and 1, respectively. Then (3) and (4) are equivalent to

$$\begin{aligned}&(-1)^d \cdot \{\mathbb{P}(Y \in B, D = d | Z = z_{k+1}) - \mathbb{P}(Y \in B, D = d | Z = z_k)\} \leq 0 \\ &\text{and } \mathbb{P}(D \in C | Z = z_{k+1}) - \mathbb{P}(D \in C | Z = z_k) \leq 0 \end{aligned} \tag{10}$$

for all k with $1 \leq k \leq K-1$, all closed intervals B in $\mathbb{R}$, each $d \in \{0, 1\}$, and all $C = (-\infty, c]$ with $c \in \mathbb{R}$. Here, (3) and (4) originally require (10) to hold for all Borel sets B. Similar to Lemma B.7 of Kitagawa (2015), we can show (by applying Lemma C1 of Andrews and Shi (2013)) that (10) holding for all closed intervals B is equivalent to (10) holding for all Borel sets B.

By definition, for all $B, C \in \mathcal{B}_{\mathbb{R}}$ and all k with $1 \leq k \leq K$, $\mathbb{P}(Y \in B, D \in C | Z = z_k) = \mathbb{P}(Y \in B, D \in C, Z = z_k) / \mathbb{P}(Z = z_k)$. We now define function spaces

$$\begin{aligned}
&\mathcal{G}_K = \left\{1_{\mathbb{R}\times\mathbb{R}\times\{z_k\}} : k = 1, 2, \ldots, K\right\}, \\
&\mathcal{G} = \left\{\left(1_{\mathbb{R}\times\mathbb{R}\times\{z_k\}}, 1_{\mathbb{R}\times\mathbb{R}\times\{z_{k+1}\}}\right) : k = 1, 2, \ldots, K-1\right\}, \\
&\mathcal{H}_1 = \left\{(-1)^d \cdot 1_{B\times\{d\}\times\mathbb{R}} : B \text{ is a closed interval in } \mathbb{R}, d \in \{0, 1\}\right\}, \\
&\bar{\mathcal{H}}_1 = \left\{(-1)^d \cdot 1_{B\times\{d\}\times\mathbb{R}} : B \text{ is a closed, open, or half-closed interval in } \mathbb{R}, d \in \{0, 1\}\right\}, \\
&\mathcal{H}_2 = \left\{1_{\mathbb{R}\times C\times\mathbb{R}} : C = (-\infty, c], c \in \mathbb{R}\right\}, \\
&\bar{\mathcal{H}}_2 = \left\{1_{\mathbb{R}\times C\times\mathbb{R}} : C = (-\infty, c] \text{ or } C = (-\infty, c), c \in \mathbb{R}\right\}, \\
&\mathcal{H} = \mathcal{H}_1 \cup \mathcal{H}_2, \text{ and } \bar{\mathcal{H}} = \bar{\mathcal{H}}_1 \cup \bar{\mathcal{H}}_2.
\end{aligned} \tag{11}$$

Let $\mathcal{P}$ denote the set of probability measures on $(\mathbb{R}^3, \mathcal{B}_{\mathbb{R}^3})$. We use an i.i.d. sample $\{(Y_i, D_i, Z_i)\}_{i=1}^n$ which is distributed according to some probability distribution Q in $\mathcal{P}$, that is, that the measure $Q(G) = \mathbb{P}((Y_i, D_i, Z_i) \in G)$ for all $G \in \mathcal{B}_{\mathbb{R}^3}$, to construct a test for the testable implication given in (3) and (4) (or in (10)). For every $Q \in \mathcal{P}$ and every measurable function v, by an abuse of notation we define

$$Q(v) = \int v \, \mathrm{d}Q. \tag{12}$$

Define, by convention (see, for example, Folland (1999, p. 45)), that

$$0 \cdot \infty = 0. \tag{13}$$

For each $Q \in \mathcal{P}$, the closure of $\mathcal{H}$ in $L^2(Q)$ is equal to $\bar{\mathcal{H}}$ (Lemma C.1). For every $Q \in \mathcal{P}$ and every $(h, g) \in \bar{\mathcal{H}} \times \mathcal{G}$ with $g = (g_1, g_2)$, define

$$\phi_Q(h, g) = \frac{Q(h \cdot g_2)}{Q(g_2)} - \frac{Q(h \cdot g_1)}{Q(g_1)}. \tag{14}$$

With (13), ϕ_Q is always well defined. Then the null hypothesis equivalent to (10) is

$$H_0 : \sup_{(h,g) \in \mathcal{H} \times \mathcal{G}} \phi_Q(h, g) \leq 0 \tag{15}$$

if the underlying distribution of the data is Q. Since $Q(v)$ is continuous on $L^2(Q)$, (15) is equivalent to $\sup_{(h,g) \in \bar{\mathcal{H}} \times \mathcal{G}} \phi_Q(h, g) \leq 0$. The alternative hypothesis is naturally set to

$$H_1 : \sup_{(h,g) \in \mathcal{H} \times \mathcal{G}} \phi_Q(h, g) > 0.$$

Define the sample analogue of ϕ_Q by

$$\hat{\phi}_Q(h, g) = \frac{\hat{Q}(h \cdot g_2)}{\hat{Q}(g_2)} - \frac{\hat{Q}(h \cdot g_1)}{\hat{Q}(g_1)},$$

where $\hat{Q}$ denotes the empirical probability measure of Q such that for every measurable function v,

$$\hat{Q}(v) = \frac{1}{n} \sum_{i=1}^{n} v(Y_i, D_i, Z_i), \tag{16}$$

and $\{(Y_i, D_i, Z_i)\}_{i=1}^{n}$ is the i.i.d. sample distributed according to Q.

The goal of this section is to construct a test for the H_0 in (15). To evaluate the ability of the test to provide size control, we consider a "local" sequence of probability distributions $\{P_n\}_{n=1}^{\infty} \subset \mathcal{P}$ under which the testable implication is true and P_n converges to some probability measure $P \in \mathcal{P}$. We introduce the next two assumptions to formalize the above settings.

Assumption 3.1 *$\{(Y_i, D_i, Z_i)\}_{i=1}^n$ is an i.i.d. data set distributed according to probability distribution P_n for each n, where D_i and Z_i are discrete variables with support $\mathcal{D}$ and $\mathcal{Z}$, respectively.*

Assumption 3.2 *There is a probability measure $P \in \mathcal{P}$ such that*

$$\lim_{n\to\infty} \int \left(\sqrt{n} \left\{ \mathrm{d}P_n^{1/2} - \mathrm{d}P^{1/2} \right\} - \frac{1}{2} v_0 \mathrm{d}P^{1/2} \right)^2 = 0 \tag{17}$$

for some measurable function v_0, where $\mathrm{d}P_n^{1/2}$ and $\mathrm{d}P^{1/2}$ denote the square roots of the densities of P_n and P, respectively.

Assumptions 3.1 and 3.2 assume an i.i.d. sample whose distribution P_n is allowed to change as n increases, and to converge to some probability distribution P as defined in (3.10.10) of van der Vaart and Wellner (1996). The local analysis of Fang and Santos (2018) considered the case where the value of the underlying parameter may be close to a point at which the map involved in the test statistic is only directionally differentiable (not fully differentiable). A similar convergent distribution sequence was introduced to show the local size control of their test.[15] As will be shown later, our test statistic involves a nondifferentiable (neither fully nor directionally differentiable) map. We follow Fang and Santos (2018) and assume such a convergent distribution sequence to show the local size control of our test.

Clearly, $\mathcal{H} \times \mathcal{G} \subset L^2(P) \times (L^2(P) \times L^2(P))$. Under Assumption 3.2, define a metric ρ_P on $L^2(P) \times (L^2(P) \times L^2(P))$ by

$$\rho_P\left((h, g), (h', g')\right) = \|h - h'\|_{L^2(P)} + \|g_1 - g_1'\|_{L^2(P)} + \|g_2 - g_2'\|_{L^2(P)} \tag{18}$$

for all $(h, g), (h', g') \in L^2(P) \times (L^2(P) \times L^2(P))$ with $g = (g_1, g_2)$ and $g' = (g_1', g_2')$. By Lemma C.8, the closure of $\mathcal{H} \times \mathcal{G}$ in $L^2(P) \times (L^2(P) \times L^2(P))$ under ρ_P is equal to $\bar{\mathcal{H}} \times \mathcal{G}$, where $\bar{\mathcal{H}}$ is defined in (11). Define

$$\Lambda(Q) = \prod_{k=1}^{K} Q\left(1_{\mathbb{R}\times\mathbb{R}\times\{z_k\}}\right) \text{ for all } Q \in \mathcal{P}, \text{ and } T_n = n \cdot \prod_{k=1}^{K} \hat{P}_n\left(1_{\mathbb{R}\times\mathbb{R}\times\{z_k\}}\right),$$

where $\hat{P}_n$ is the empirical probability measure of P_n defined as in (16). Under Assumption 3.2, we mainly consider the nontrivial case where $\Lambda(P) > 0$. Also, for every $Q \in \mathcal{P}$, define

$$\sigma_Q^2(h, g) = \Lambda(Q) \cdot \left\{ \frac{Q\left(h^2 \cdot g_2\right)}{Q^2\left(g_2\right)} - \frac{Q^2\left(h \cdot g_2\right)}{Q^3\left(g_2\right)} + \frac{Q\left(h^2 \cdot g_1\right)}{Q^2\left(g_1\right)} - \frac{Q^2\left(h \cdot g_1\right)}{Q^3\left(g_1\right)} \right\} \tag{19}$$

[15] See Examples 2.1 and 2.2 of Fang and Santos (2018).

for all $(h, g) \in \bar{\mathcal{H}} \times \mathcal{G}$ with $g = (g_1, g_2)$, where $Q^m(v) = [Q(v)]^m$ for all $m \in \mathbb{N}$ and all measurable v.

Lemma 3.1 *Under Assumptions 3.1 and 3.2, $\sqrt{T_n}(\hat{\phi}_{P_n} - \phi_P) \rightsquigarrow \mathbb{G}$ for some tight*[16] *random element $\mathbb{G}$ which almost surely has a uniformly ρ_P-continuous path, and for all $(h, g) \in \bar{\mathcal{H}} \times \mathcal{G}$ with $g = (g_1, g_2)$, the variance $Var(\mathbb{G}(h, g))$ is equal to the $\sigma_P^2(h, g)$ given in (19), where*

$$\sigma_P^2(h, g) \leq 1/4 \cdot \max_{(g_1', g_2') \in \mathcal{G}} \left\{ \Lambda(P) / P(g_2') + \Lambda(P) / P(g_1') \right\} \leq 1/2 \cdot (K-1)^{-(K-1)}, \quad (20)$$

and K is the number of elements of $\mathcal{Z}$. In particular, $\sigma_P^2(h, g) \leq 1/4$ for all $(h, g) \in \bar{\mathcal{H}} \times \mathcal{G}$ when $K = 2$.

Lemma 3.1 provides the pointwise (P is fixed) asymptotic distribution of $\sqrt{T_n}(\hat{\phi}_{P_n} - \phi_P)$ as P_n converges to P under Assumption 3.2. We note that the pointwise asymptotic distribution of $\sqrt{T_n}(\hat{\phi}_{P_n} - \phi_P)$ is different from the asymptotic distribution of $\sqrt{T_n}(\hat{\phi}_{P_n} - \phi_{P_n})$ which can be obtained by Theorem 3.10.12 of van der Vaart and Wellner (1996). The weak convergence of $\sqrt{T_n}(\hat{\phi}_Q - \phi_Q)$ uniform in Q may be obtained under different assumptions following the notion of van der Vaart and Wellner (1996, p. 168). We derive the pointwise asymptotic distribution in order to obtain the null asymptotic distribution of the test statistic using the proposed extended delta method. See the discussion after Theorem 3.1. Lemma 3.1 also provides the asymptotic variance of $\sqrt{T_n}(\hat{\phi}_{P_n} - \phi_P)$, which is uniformly bounded by 1 for all $K > 1$. We used the quantity $\sqrt{T_n}$ instead of $\sqrt{n}$ to establish the asymptotic distribution in order to achieve a parameter-free bound for the asymptotic variance as shown in (20).[17] The quantity T_n is asymptotically equivalent to n in the sense that $T_n/n \to \prod_{k=1}^K \mathbb{P}(Z = z_k)$ in probability. If we use $\sqrt{n}$, the bound of the asymptotic variance may involve the underlying parameter P. In the binary instrument case where $Z \in \{0, 1\}$, we let $m_0 = \sum_{i=1}^n 1\{Z_i = 0\}$ and $m_1 = \sum_{i=1}^n 1\{Z_i = 1\}$. It then follows that $T_n = m_0 m_1 / n$ which is used in the test of Kitagawa (2015). Suppose instead $Z \in \{0, 1, 2\}$, and we let $m_z = \sum_{i=1}^n 1\{Z_i = z\}$ for $z \in \{0, 1, 2\}$. Then $T_n = m_0 m_1 m_2 / n^2$.

The bound in (20) will be useful when we construct the test statistic. By (19), for every

[16] In a metric space, tightness implies separability.

[17] In practice, when the sample size is small, it is possible that we only have a small number of observations for $Z = z_k$ for some k. In this case, we can use $\sqrt{n}$ instead of $\sqrt{T_n}$ to construct the test statistic. We then use (21) to find an empirical bound for $\hat{\sigma}_{P_n}$, and use this bound to determine the values of ξ. We could also redefine the instrument Z, in some cases, such that we have more observations for each possible value of the redefined instrument. For example, we may define the new instrument $\tilde{Z} = 1\{Z \geq z_0\}$ for some z_0. However, this will change the definitions of all types of individuals (always takers, compliers, defiers, and never takers). In this case, we need to guarantee that the instrument used in the empirical analysis and the instrument used in the test are the same. The test result for $\tilde{Z}$ may be false for Z.

$(h,g) \in \bar{\mathcal{H}} \times \mathcal{G}$ with $g = (g_1, g_2)$, define the sample analogue of $\sigma_P^2(h,g)$ by

$$\hat{\sigma}_{P_n}^2(h,g) = \frac{T_n}{n} \cdot \left\{ \frac{\hat{P}_n(h^2 \cdot g_2)}{\hat{P}_n^2(g_2)} - \frac{\hat{P}_n^2(h \cdot g_2)}{\hat{P}_n^3(g_2)} + \frac{\hat{P}_n(h^2 \cdot g_1)}{\hat{P}_n^2(g_1)} - \frac{\hat{P}_n^2(h \cdot g_1)}{\hat{P}_n^3(g_1)} \right\}. \tag{21}$$

Note that for each $h \in \bar{\mathcal{H}}$ and each $g_l \in \mathcal{G}_K$, if $\hat{P}_n(g_l) = 0$ then $\hat{P}_n(h \cdot g_l) = 0$. By (13), $\hat{\sigma}_{P_n}^2$ is well defined. Similar to (20), we can find a bound for $\hat{\sigma}_{P_n}$ for every finite sample. It can be shown that for all (h,g),

$$\hat{\sigma}_{P_n}^2(h,g) \le 1/4 \cdot \max_{(g_1',g_2') \in \mathcal{G}} \left\{ (T_n/n)/\hat{P}_n(g_2') + (T_n/n)/\hat{P}_n(g_1') \right\} \le 1/2 \cdot (K-1)^{-(K-1)}. \tag{22}$$

Clearly, the bounds for σ_P and $\hat{\sigma}_{P_n}$ will decrease as K increases.

We may extend the idea of Kitagawa (2015) and construct the test statistic to be

$$\sup_{(h,g) \in \mathcal{H} \times \mathcal{G}} \frac{\sqrt{T_n} \hat{\phi}_{P_n}(h,g)}{\max\{\xi, \hat{\sigma}_{P_n}(h,g)\}} \tag{23}$$

for some positive number (trimming parameter) ξ. Here, ξ plays two roles: (1) Since $\hat{\sigma}_{P_n}$ can be zero, ξ bounds the denominator away from zero; (2) as shown in the Monte Carlo studies of Kitagawa (2015) and the present paper, different values of ξ, from small (close to 0) to large (close to 1), may lead to different powers of the test for the same data generating process (DGP), which could be close to 0. Kitagawa (2015) suggests that if there is no prior knowledge available about a likely alternative, the default choice of ξ could be set to 0.07 according to the simulation studies for the binary treatment and binary instrument case. They also suggest that users report test results using different values of ξ. This paper constructs the test statistic in a way that, loosely speaking, computes the weighted average of the test statistics in (23) over ξ.[18] If we put all the weight on one particular value of ξ, the test statistic degenerates to the test statistic in (23).

Let Ξ be a predetermined closed subset of $[0,1]$ such that $0 \notin \Xi$. The set Ξ contains all the values of ξ used for constructing the test statistic. Only one of the values greater than (or equal to) the bound in Lemma 3.1, say 1, needs to be included in Ξ. The test statistic in (23) reduces to the unweighted KS statistic when $\xi = 1$. Let ν be a positive measure on Ξ.

Assumption 3.3 *The measure ν satisfies that $0 < \nu(\Xi) < \infty$ and $S_n \in L^1(\nu)$ for all $\omega \in \Omega$*

[18] In this way, we can avoid repeating the test using the same data set but different values of ξ and making a decision based on all these results. The potential issue of multiple comparisons can be prevented accordingly.

and all n with

$$S_n(\xi) = \sup_{(h,g)\in\mathcal{H}\times\mathcal{G}} \frac{\hat{\phi}_{P_n}(h,g)}{\max\{\xi, \hat{\sigma}_{P_n}(h,g)\}}.$$

Now we set the test statistic to

$$TS_n = \int_{\Xi} \sup_{(h,g)\in\mathcal{H}\times\mathcal{G}} \frac{\sqrt{T_n}\hat{\phi}_{P_n}(h,g)}{\max\{\xi, \hat{\sigma}_{P_n}(h,g)\}} \mathrm{d}\nu(\xi). \tag{24}$$

The measure ν could be a Dirac measure centered at some fixed $\xi \in \Xi$. This is equivalent to using a particular value for the trimming parameter to construct the test statistic as in (23). Or ν could be a discrete or continuous probability measure that assigns probabilities to the elements of Ξ. This is equivalent to using a weighted average of the test statistics in (23) over ξ. By using (24), we take into account the fact that the values of ξ may influence the power of the test, and we can also avoid the multiple testing issue. See the discussion in Section 4 about the computational simplification of the test statistic in (24). Define

$$\Psi_{\mathcal{H}\times\mathcal{G}} = \{(h,g) \in \mathcal{H}\times\mathcal{G} : \phi_P(h,g) = 0\} \text{ and } \Psi_{\bar{\mathcal{H}}\times\mathcal{G}} = \left\{(h,g) \in \bar{\mathcal{H}}\times\mathcal{G} : \phi_P(h,g) = 0\right\}. \tag{25}$$

Since $1_{\{a\}\times\{0\}\times\mathbb{R}}, -1_{\{a\}\times\{1\}\times\mathbb{R}} \in \mathcal{H}$ for all $a \in \mathbb{R}$, $\Psi_{\mathcal{H}\times\mathcal{G}}$ and $\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}$ are not empty.

In the following theorem, we establish the asymptotic distribution of the test statistic under null. We note that the L^r ($r \in \mathbb{N}$) spaces play an important role in deriving this asymptotic distribution. For example, we show that $\bar{\mathcal{H}}$ is compact in $L^2(Q)$ for every $Q \in \mathcal{P}$ and $\bar{\mathcal{H}} \times \mathcal{G}$ is compact in $L^2(P) \times (L^2(P) \times L^2(P))$ under ρ_P (constructed based on the L^2 norm). We obtain the Glivenko–Cantelli and the Donsker results using the L^1 and the L^2 norms. We also show that the weak limit $\mathbb{G}$ of $\sqrt{T_n}(\hat{\phi}_{P_n} - \phi_P)$ in Lemma 3.1 has a continuous path under ρ_P.[19] The weak convergence in Theorem 3.1 is established by using these fundamental results.

Theorem 3.1 *Suppose Assumptions 3.1, 3.2, and 3.3 hold. If the H_0 in (15) is true with $Q = P_n$ for all n, then*

$$TS_n \rightsquigarrow \int_{\Xi} \sup_{(h,g)\in\Psi_{\mathcal{H}\times\mathcal{G}}} \frac{\mathbb{G}(h,g)}{\max\{\xi, \sigma_P(h,g)\}} \mathrm{d}\nu(\xi), \tag{26}$$

where $\mathbb{G}$ is as in Lemma 3.1.

Theorem 3.1 provides the pointwise (P is fixed) asymptotic distribution of the test statistic

[19] See Appendix C for more detailed results.

if the H_0 in (15) is true with $Q = P_n$ for all n.[20] To find this asymptotic distribution, we employed the pointwise weak convergence in Lemma 3.1 and the extended delta method provided in Appendix A. Because of a nondifferentiability issue, the existing delta methods fail to work in establishing the weak convergence in (26). In Appendix A, we provide an extended continuous mapping theorem and an extended delta method elaborated by Theorems A.1 and A.2, respectively, to deal with this technical issue. See further discussion in Remark C.3. Theorem A.1 can be viewed as an extension of Theorem 1.11.1 of van der Vaart and Wellner (1996), and Theorem A.2 can be viewed as an extension of Theorem 3.9.5 of van der Vaart and Wellner (1996) and of Theorem 2.1 of Fang and Santos (2018). For simplicity of notation, we let

$$\mathbb{T} = \int_{\Xi} \sup_{(h,g)\in\Psi_{\mathcal{H}\times\mathcal{G}}} \frac{\mathbb{G}(h,g)}{\max\{\xi, \sigma_P(h,g)\}} \mathrm{d}\nu(\xi) \text{ and } \mathbb{T}_0 = \int_{\Xi} \sup_{(h,g)\in\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}} \frac{\mathbb{G}_0(h,g)}{\max\{\xi, \sigma_P(h,g)\}} \mathrm{d}\nu(\xi),$$

where $\mathbb{G}_0$ is some random element such that $\mathbb{G} = \mathbb{G}_0 + \Lambda(P)^{1/2}\mathcal{L}'_P(Q_0)$,[21] where $Q_0(v) = P(vv_0)$ for all suitable v, and for all $(h,g) \in \bar{\mathcal{H}} \times \mathcal{G}$,

$$\mathcal{L}'_P(Q_0)(h,g) = \frac{Q_0(h\cdot g_2)P(g_2) - P(h\cdot g_2)Q_0(g_2)}{P^2(g_2)} - \frac{Q_0(h\cdot g_1)P(g_1) - P(h\cdot g_1)Q_0(g_1)}{P^2(g_1)}.$$

It can be shown that $\mathcal{L}'_P(Q_0) \le 0$ on $\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}$ under H_0, and thus $\mathbb{G} \le \mathbb{G}_0$. Following the proof of Theorem 3.1, we can show that

$$\mathbb{T} = \int_{\Xi} \sup_{(h,g)\in\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}} \frac{\mathbb{G}(h,g)}{\max\{\xi, \sigma_P(h,g)\}} \mathrm{d}\nu(\xi).$$

It then follows that $\mathbb{T} \le \mathbb{T}_0$. When P_n is fixed at some P for all n, then $v_0 = 0$ and $\mathbb{G}_0 = \mathbb{G}$.

3.1 Bootstrap-Based Inference

It was shown that the asymptotic distribution in (26) involves the set $\Psi_{\mathcal{H}\times\mathcal{G}}$ that depends on the underlying probability measure P. Therefore, we need to find a "valid" estimator $\widehat{\Psi_{\mathcal{H}\times\mathcal{G}}}$ for $\Psi_{\mathcal{H}\times\mathcal{G}}$ in order to consistently approximate the asymptotic distribution. By the definition of $\Psi_{\mathcal{H}\times\mathcal{G}}$ in (25), we construct $\widehat{\Psi_{\mathcal{H}\times\mathcal{G}}}$ by

$$\widehat{\Psi_{\mathcal{H}\times\mathcal{G}}} = \left\{(h,g) \in \mathcal{H}\times\mathcal{G} : \sqrt{T_n}\left|\frac{\hat{\phi}_{P_n}(h,g)}{\max\{\xi_0, \hat{\sigma}_{P_n}(h,g)\}}\right| \le \tau_n\right\} \tag{27}$$

[20] More precisely, the weak convergence in (26) is under P_n.
[21] See more details in the proof of Theorem 3.2.

with $\tau_n \to \infty$ and $\tau_n/\sqrt{n} \to 0$ as $n \to \infty$, where ξ_0 is a small positive number. We suggest using $\xi_0 = 0.001$ in practice.[22] This is a method similar to that which is used in Beare and Shi (2019) and Sun and Beare (2021) to estimate contact sets in independent contexts. See Linton et al. (2010) and Lee et al. (2013) for further discussion of estimation of contact sets. Each (h, g) is included in $\widehat{\Psi_{\mathcal{H}\times\mathcal{G}}}$ if $\sqrt{T_n}|\hat{\phi}_{P_n}(h, g)|$ is no more than τ_n estimated standard deviations from zero. As mentioned by Sun and Beare (2021), we can effectively use pointwise confidence intervals to select points in this way.

3.1.1 Test Procedure

We implement the test in the following sequence of steps:

(1) Obtain the bootstrap sample $\{(\hat{Y}_i, \hat{D}_i, \hat{Z}_i)\}_{i=1}^n$ drawn independently with replacement from the sample $\{(Y_i, D_i, Z_i)\}_{i=1}^n$.

(2) Calculate the bootstrap version of $\hat{\phi}_{P_n}$ by

$$\hat{\phi}_{P_n}^B(h, g) = \frac{\hat{P}_n^B(h \cdot g_2)}{\hat{P}_n^B(g_2)} - \frac{\hat{P}_n^B(h \cdot g_1)}{\hat{P}_n^B(g_1)}, \tag{28}$$

let $T_n^B = n \cdot \prod_{k=1}^K \hat{P}_n^B\left(1_{\mathbb{R}\times\mathbb{R}\times\{z_k\}}\right)$, and calculate the bootstrap version of $\hat{\sigma}_{P_n}$ by

$$\hat{\sigma}_{P_n}^B(h, g) = \sqrt{\frac{T_n^B}{n}} \cdot \sqrt{\frac{\hat{P}_n^B(h^2 \cdot g_2)}{\hat{P}_n^B(g_2)^2} - \frac{\hat{P}_n^B(h \cdot g_2)^2}{\hat{P}_n^B(g_2)^3} + \frac{\hat{P}_n^B(h^2 \cdot g_1)}{\hat{P}_n^B(g_1)^2} - \frac{\hat{P}_n^B(h \cdot g_1)^2}{\hat{P}_n^B(g_1)^3}} \tag{29}$$

for all $(h, g) \in \bar{\mathcal{H}} \times \mathcal{G}$ with $g = (g_1, g_2)$, where $\hat{P}_n^B(v) = n^{-1}\sum_{i=1}^n v(\hat{Y}_i, \hat{D}_i, \hat{Z}_i)$ for all measurable v. We note that (22) also provides a bound for $\hat{\sigma}_{P_n}^B$.

(3) Calculate the bootstrap version of the test statistic by

$$TS_n^B = \int_\Xi \sup_{(h,g)\in\widehat{\Psi_{\mathcal{H}\times\mathcal{G}}}} \frac{\sqrt{T_n^B}(\hat{\phi}_{P_n}^B(h, g) - \hat{\phi}_{P_n}(h, g))}{\max\{\xi, \hat{\sigma}_{P_n}^B(h, g)\}} \, \mathrm{d}\nu(\xi). \tag{30}$$

Since the asymptotic distribution in (26) involves a nonlinear map, the bootstrap test statistic in (30) was constructed following the idea of Fang and Santos (2018). The nonlinearity of the map may cause inconsistencies in the "standard" bootstrap approximation. See Dümbgen (1993), Andrews (2000), and Fang and Santos (2018)

[22] It can be shown that $\widehat{\Psi_{\mathcal{H}\times\mathcal{G}}}$ can also be used to approximate the asymptotic distribution when $\mathcal{D} = \{d_1, d_2, \ldots\}$. See (C.49).

for details. Because of the denominator $\max\{\xi, \hat{\sigma}^B_{P_n}(h,g)\}$, our approach is an extension of that of Fang and Santos (2018). The calculation of (30) can be simplified in practice as discussed in Section 4 for (24).

(4) Repeat steps (1), (2), and (3) n_B times independently, for (say) $n_B = 1000$. Given the nominal significance level α, calculate the bootstrap critical value $\hat{c}_{1-\alpha}$ by

$$\hat{c}_{1-\alpha} = \inf\left\{c : \mathbb{P}\left(TS^B_n \le c \middle| \{(Y_i, D_i, Z_i)\}^n_{i=1}\right) \ge 1-\alpha\right\}. \tag{31}$$

In practice, we approximate $\hat{c}_{1-\alpha}$ by computing the $1-\alpha$ quantile of the n_B independently generated bootstrap statistics, with n_B chosen as large as is computationally convenient.

(5) The decision rule for the test is: Reject H_0 if $TS_n > \hat{c}_{1-\alpha}$.

The following theorem presents the asymptotic properties of the proposed test. Under Assumption 3.2, Theorem 3.2(i) provides the local size control of the test. As discussed in Fang and Santos (2018), the asymptotic distribution of the test statistic may discontinuously depend on the parameter of interest, if the map involved in the test statistic is not fully differentiable. However, the finite sample distribution of the test statistic often continuously depends on the parameter of interest. Imbens and Manski (2004) emphasize that this discrepancy may cause poor finite sample properties of the test. As suggested by Fang and Santos (2018), a local analysis can help better approximate the finite sample properties of the test when the parameter of interest is close to a point at which the map is not fully differentiable. Our test statistic involves a nondifferentiable map, and Theorem 3.2(i) provides evidence for the good finite sample size property of the test.

Theorem 3.2 *Suppose Assumptions 3.1, 3.2, and 3.3 hold.*

(i) If the H_0 in (15) is true with $Q = P_n$ for all n and the CDF of $\mathbb{T}_0$ is increasing and continuous at its $1-\alpha$ quantile $c_{1-\alpha}$, then $\lim_{n\to\infty} \mathbb{P}(TS_n > \hat{c}_{1-\alpha}) \le \alpha$. If, in addition, $P_n = P$ for all large n, then $\lim_{n\to\infty} \mathbb{P}(TS_n > \hat{c}_{1-\alpha}) = \alpha$.

(ii) If the H_0 in (15) is false with $Q = P$ and $P_n = P$ for all large n, then $\lim_{n\to\infty} \mathbb{P}(TS_n > \hat{c}_{1-\alpha}) = 1$.

It is implied by Theorem 11.1 of Davydov et al. (1998) that in (i) of Theorem 3.2, the CDF of $\mathbb{T}_0$ is differentiable and has a positive derivative everywhere except at countably many points in its support, provided that $\mathbb{T}_0 \ne 0$. If $\mathbb{T}_0 = 0$ at null configurations, our test statistic converges to zero in probability and so does the critical value. Theorem 3.2 does not show clearly how the rejection rate of the test will behave asymptotically in this case.

As discussed in Sun and Beare (2021), this is a common theoretical limitation for irregular testing problems. Tests based on the machinery of Fang and Santos (2018), and also those based on generalized moment selection (Andrews and Soares, 2010; Andrews and Shi, 2013), may encounter this issue. One practical resolution is to replace the bootstrap critical value $\hat{c}_{1-\alpha}$ with $\max\{\hat{c}_{1-\alpha}, \eta\}$ or $\hat{c}_{1-\alpha} + \eta$, where η is some small positive constant. See, for instance, Donald and Hsu (2016, p. 13). Simulation results in Table 5 showed that the empirical rejection rates of our test with $\eta = 0$ ($\tau_n = 2$) are well controlled by the nominal significance level when $\mathbb{T}_0 = 0$ under null configurations.[23]

Theorem 3.2(i) shows that the test is locally size controlled for every convergent distribution sequence satisfying the null. The convergent probability distributions $\{P_n\}$ depend on n, that is, the underlying distribution P_n of the data can be different for every n. As $n \to \infty$, P_n (satisfying the null) converges to P under Assumption 3.2. Theorem 3.1 provides the pointwise (P is fixed) asymptotic distribution of the test statistic TS_n for this convergent sequence of probability distributions $P_n \to P$. With this pointwise asymptotic distribution, we then obtain the local size control along such a probability distribution sequence: $\lim_{n\to\infty} \mathbb{P}(TS_n > \hat{c}_{1-\alpha}) \leq \alpha$. When both D and Z are binary, Kitagawa (2015) and the present paper consider testing the same null and alternative hypotheses. Kitagawa (2015) obtains the uniform size control for their test under different conditions. That is, $\limsup_{n\to\infty} \sup_{Q\in\mathcal{P}_0} \mathbb{P}(TS_n > \hat{c}^K_{1-\alpha}) \leq \alpha$, where $\mathcal{P}_0$ denotes the set of probability distributions in $\mathcal{P}$ that satisfy H_0, and the superscript "K" denotes the critical value of Kitagawa (2015) (the test statistic of Kitagawa (2015) is equivalent to that in the present paper when both D and Z are binary). Since Theorem 3.2(i) assumes $P_n \in \mathcal{P}_0$, clearly we have that for every P_n,

$$\mathbb{P}(TS_n > \hat{c}^K_{1-\alpha}) \leq \sup_{Q\in\mathcal{P}_0} \mathbb{P}(TS_n > \hat{c}^K_{1-\alpha}),$$

which indicates that the uniform size control of Kitagawa (2015) implies local size control. In general, without additional assumptions, the local size control of the proposed test does not directly imply the uniform size control over the class of data generating processes in the null.

3.2 Binary Treatment and Binary Instrument: Power Improvement

In this section, we consider the special case where the treatment D and the instrument Z are both binary. We show how to achieve power improvement over the test of Kitagawa (2015) based on the results of Kitagawa (2015) and those from Section 3.1. As shown in Section 2, the null hypothesis for the testable implications consists of a set of inequalities.

[23] In Section 4, the value of τ_n is chosen to be 2.

Kitagawa (2015) used an upper bound on the asymptotic distribution of the test statistic under null to construct the bootstrap critical value. The upper bound is identical to the asymptotic distribution when all the inequalities in the null are binding. Therefore, their test could be conservative. The present paper establishes the asymptotic distribution of the test statistic under null. We then construct the critical value based on this asymptotic distribution, rather than on an upper bound, and therefore improve the power of the test.

Let $z_1 = 0$, $z_2 = 1$, $d_1 = 0$, and $d_2 = 1$. The test statistic in (24) is now numerically equal to the one constructed by Kitagawa (2015) if we let ν be a Dirac measure. Recall that the instrument is allowed to be multivalued under the constructions in Section 3.[24]

We consider a simple case where $P_n = P$ for all n and the H_0 in (15) is true with $Q = P$. As introduced in Section 2, we follow Kitagawa (2015) and define probability measures

$$P_1(B, C) = \mathbb{P}(Y \in B, D \in C | Z = 1) \text{ and } P_0(B, C) = \mathbb{P}(Y \in B, D \in C | Z = 0)$$

for all $B, C \in \mathcal{B}_{\mathbb{R}}$. Now we define

$$\mathcal{F}_b = \left\{ (-1)^d \cdot 1_{B \times \{d\}} : B \text{ is a closed interval}, d \in \{0, 1\} \right\},$$

and write $P_d(f) = \int f \, \mathrm{d}P_d$ for all measurable f and each $d \in \{0, 1\}$. Kitagawa (2015) showed that their critical value converged to the $1 - \alpha$ quantile of the distribution $\sup_{f \in \mathcal{F}_b} \mathbb{G}_H(f)/(\xi \vee \sigma_H(f))$, where $H = \lambda P_1 + (1 - \lambda) P_0$, $\lambda = \mathbb{P}(Z = 1)$, $\mathbb{G}_H$ is an H-Brownian bridge, and $\sigma_H(f)$ is the standard deviation of $\mathbb{G}_H(f)$, that is, $\sigma_H^2(f) = H(f^2) - H^2(f)$. Let $\mathcal{F}_b^* = \{f \in \mathcal{F}_b : P_0(f) = P_1(f)\}$. Then it is easy to show that $H(f) = P_0(f) = P_1(f)$ for all $f \in \mathcal{F}_b^*$. Let ν be a Dirac measure centered at some ξ. It can be shown that

$$\sup_{f \in \mathcal{F}_b} \frac{\mathbb{G}_H(f)}{\xi \vee \sigma_H(f)} \geq \sup_{f \in \mathcal{F}_b^*} \frac{\mathbb{G}_H(f)}{\xi \vee \sigma_H(f)} \stackrel{L}{=} \mathbb{T}, \tag{32}$$

where $\mathbb{T}$ is the asymptotic distribution of the test statistic in (26) and "$\stackrel{L}{=}$" means equivalence in distribution.

Kitagawa (2015) constructed a pooled-data bootstrap approximation for the Gaussian process $\mathbb{G}_H/(\xi \vee \sigma_H)$, denoted by $\mathbb{G}_H^B/(\xi \vee \sigma_H^B)$, and then computed the bootstrap test statistic by $\sup_{f \in \mathcal{F}_b} \mathbb{G}_H^B(f)/(\xi \vee \sigma_H^B(f))$. This bootstrap statistic approximates the distri-

[24]For the case where the treatment is binary and the instrument is multivalued, Kitagawa (2015) constructed the test statistic by first computing the normalized differences of two empirical probability measures between neighboring pairs of values of instruments (ordered according to the propensity score), and then taking the maximum value of all these differences. Since these differences can be mutually correlated, it would not be straightforward to obtain the asymptotic distribution of their test statistic and approximate its null distribution by bootstrap.

bution of $\sup_{f\in\mathcal{F}_b}\mathbb{G}_H(f)/(\xi\vee\sigma_H(f))$. For the case where D and Z are both binary, we suggest modifying the test in Section 3.1 to achieve power improvement over the test of Kitagawa (2015).[25] Specifically, we first estimate $\mathcal{F}_b^*$ by a subset of $\mathcal{F}_b$, denoted by $\widehat{\mathcal{F}_b^*}$, in a way similar to (27). Then we follow the bootstrap approach of Kitagawa (2015) to construct $\mathbb{G}_H^B$ and σ_H^B, and construct the bootstrap test statistic by $\sup_{f\in\widehat{\mathcal{F}_b^*}}\mathbb{G}_H^B(f)/(\xi\vee\sigma_H^B(f))$. Clearly, the proposed critical value is always smaller than that of Kitagawa (2015) because $\widehat{\mathcal{F}_b^*}\subset\mathcal{F}_b$. It can also be shown that our critical value converges to the $1-\alpha$ quantile of $\sup_{f\in\mathcal{F}_b^*}\mathbb{G}_H(f)/(\xi\vee\sigma_H(f))$ (equivalently, $\mathbb{T}$) under H_0. Since the test statistic in (24) is numerically equivalent to that of Kitagawa (2015), this shows that the power of the test can be improved by the use of our approach. This improvement is against all alternatives according to the construction of the critical value. See the simulation evidence in Appendix E.4.

The test of Mourifié and Wan (2017) for the inequalities in (2) employed the intersection bounds framework of Chernozhukov et al. (2013). As shown in Proposition 1 of Mourifié and Wan (2017),[26] the limiting rejection rate under null is equal to the nominal significance level α only when all the inequalities in the null are binding, and below the nominal level elsewhere in the null. This result is similar to that of Kitagawa (2015), because only when all the inequalities are binding, the (contact) set $\mathcal{F}_b^*$ is equal to $\mathcal{F}_b$. If we are at a point in the null where the inequalities are not all binding, then the tests of Kitagawa (2015) and Mourifié and Wan (2017) would have limiting rejection rates below the nominal level, and thus lack power against nearby points in the alternative. Theorem 3.2 in the present paper shows that the proposed test can achieve the nominal level over a larger region in the null, where the inequalities in the testable implication could not all be binding, thereby improving power.

3.3 Unordered Treatment

With testable implication (5), we define the function space

$$\mathcal{H}\times\mathcal{G}=\left\{\left(1_{B\times\{d\}\times\mathbb{R}},\left(1_{\mathbb{R}\times\mathbb{R}\times\{z\}},1_{\mathbb{R}\times\mathbb{R}\times\{z'\}}\right)\right):B\text{ is a closed interval, }\left(d,z,z'\right)\in\mathcal{C}\right\}. \quad (33)$$

For every probability measure Q with (12), we define ϕ_Q by $\phi_Q(h,g)=Q(h\cdot g_2)/Q(g_2)-Q(h\cdot g_1)/Q(g_1)$ for every $(h,g)\in\mathcal{H}\times\mathcal{G}$ with $g=(g_1,g_2)$. Testable implication (5) is equivalent to the H_0 in

$$H_0:\sup_{(h,g)\in\mathcal{H}\times\mathcal{G}}\phi_Q(h,g)\le 0\text{ and }H_1:\sup_{(h,g)\in\mathcal{H}\times\mathcal{G}}\phi_Q(h,g)>0$$

[25] The modification may also be applied to the case where D is multivalued and Z is binary.
[26] See also Theorem 6 of Chernozhukov et al. (2013).

if Q is the underlying probability distribution of the data. Then we can follow the test procedure in Section 3.1.1 to conduct the test with the function space $\mathcal{H} \times \mathcal{G}$ defined in (33).

4 Simulation Evidence

We first designed Monte Carlo simulations for the case where D and Z are both multivalued random variables such that $D \in \{0, 1, 2\}$ and $Z \in \{0, 1, 2\}$. Additional Monte Carlo studies can be found in Appendix E. Each simulation consisted of 1000 Monte Carlo iterations and 1000 bootstrap iterations. To expedite the simulation, we employed the warp-speed method of Giacomini et al. (2013). The nominal significance level α was set to 0.05. As shown in (20) and (22), σ_P^2 and $\hat{\sigma}_{P_n}^2$ are bounded by $(1/2)\cdot(K-1)^{-(K-1)}$, where $K = 3$ in our setting. The simulations constructed in this section are similar to those in Kitagawa (2015). In each simulation, the measure ν was set to be a Dirac measure δ_ξ centered at one of the following values of ξ: 0.07, 0.1, 0.13, 0.16, 0.19, 0.22, 0.25, 0.28, 0.3, and 1, or to be a probability measure $\bar{\nu}_\xi$ that assigns equal probabilities (weights) to the values of ξ listed above. Four values of ξ were used in the simulations of Kitagawa (2015): 0.07, 0.22, 0.3, and 1, where $0.07 \approx \sqrt{0.005(1-0.005)}$, $0.22 \approx \sqrt{0.05(1-0.05)}$, and $0.3 = \sqrt{0.1(1-0.1)}$. As shown in (21), for every $(h, g) \in \bar{\mathcal{H}} \times \mathcal{G}$ with $g = (g_1, g_2)$,

$$
\begin{aligned}
\hat{\sigma}_{P_n}^2(h,g) &= \frac{T_n}{n}\cdot\left\{\frac{\hat{P}_n\left(h^2\cdot g_2\right)}{\hat{P}_n^2\left(g_2\right)} - \frac{\hat{P}_n^2\left(h\cdot g_2\right)}{\hat{P}_n^3\left(g_2\right)} + \frac{\hat{P}_n\left(h^2\cdot g_1\right)}{\hat{P}_n^2\left(g_1\right)} - \frac{\hat{P}_n^2\left(h\cdot g_1\right)}{\hat{P}_n^3\left(g_1\right)}\right\} \\
&= \frac{\prod_{k=1}^K \hat{P}_n\left(1_{\mathbb{R}\times\mathbb{R}\times\{z_k\}}\right)}{\hat{P}_n\left(g_2\right)}\frac{\hat{P}_n\left(h^2\cdot g_2\right)}{\hat{P}_n\left(g_2\right)}\left\{1-\frac{\hat{P}_n\left(h^2\cdot g_2\right)}{\hat{P}_n\left(g_2\right)}\right\} \\
&\quad + \frac{\prod_{k=1}^K \hat{P}_n\left(1_{\mathbb{R}\times\mathbb{R}\times\{z_k\}}\right)}{\hat{P}_n\left(g_1\right)}\frac{\hat{P}_n\left(h^2\cdot g_1\right)}{\hat{P}_n\left(g_1\right)}\left\{1-\frac{\hat{P}_n\left(h^2\cdot g_1\right)}{\hat{P}_n\left(g_1\right)}\right\}.
\end{aligned}
$$

The values of $\xi \in \{0.07, 0.22, 0.3\}$ take the form of $\sqrt{\pi(1-\pi)}$ where $\pi \in \{0.005, 0.05, 0.1\}$. As discussed in Kitagawa (2015), π can be interpreted as that if both $\hat{P}_n\left(h^2\cdot g_1\right)/\hat{P}_n\left(g_1\right)$ and $\hat{P}_n\left(h^2\cdot g_2\right)/\hat{P}_n\left(g_2\right)$ are less than π, then the weight becomes the inverse of ξ instead of the inverse of the estimated standard deviation. As π gets larger, less weight is put on $\hat{\phi}_{P_n}$ for smaller probability events, and vice versa. In the following simulations, we chose the values of ξ following the choice of Kitagawa (2015). In empirical practice, application-based simulations can be applied to choose Ξ and ν, which is illustrated in Section E.5.

When calculating the supremum in the test statistic TS_n in (24), we followed the numerical computation approach used by Kitagawa (2015). Specifically, we calculated the supremum using only the closed intervals B with the values of $\{Y_i\}_{i=1}^n$ observed in the data as the endpoints, that is, $B = [a, b]$ with $a, b \in \{Y_1, Y_2, \ldots, Y_n\}$ and $a \le b$. It is not hard to show that the test statistic calculated in this way is equal to that in (24). We also used such closed intervals to calculate the bootstrap test statistic TS_n^B in (30). From all such intervals, we found those that satisfy the inequality in (27) and used them to calculate the supremum of $\sqrt{T_n^B}(\hat{\phi}_{P_n}^B - \hat{\phi}_{P_n})/\max\{\xi, \hat{\sigma}_{P_n}^B\}$ for each ξ listed above.

4.1 Size Control and Tuning Parameter Selection

The first set of simulations was designed to investigate the size of the test and the selection of the tuning parameter. As shown in (27), the estimate $\widehat{\Psi_{\mathcal{H}\times\mathcal{G}}}$ involves a tuning parameter τ_n with $\tau_n \to \infty$ and $\tau_n/\sqrt{n} \to 0$ as $n \to \infty$. In practice, we need to use a particular value of τ_n for each sample size n. For this set of simulations, we set n to 3000 and τ_n to 0.1, 0.5, 1, 2, 3, 4, and ∞. For $\tau_n = \infty$, $\widehat{\Psi_{\mathcal{H}\times\mathcal{G}}} = \mathcal{H} \times \mathcal{G}$ and the test is conservative. We compared the rejection rates obtained using each of these values of τ_n and decided which value would be a good option for sample sizes close to 3000. We let $U \sim \mathrm{Unif}(0,1)$, $V \sim \mathrm{Unif}(0,1)$, $N_0 \sim \mathrm{N}(0,1)$, $N_1 \sim \mathrm{N}(1,1)$, $N_2 \sim \mathrm{N}(2,1)$, $Z = 2 \times 1\{U \le 0.5\} + 1\{0.5 < U \le 0.7\}$ ($\mathbb{P}(Z = 2) = 0.5$), $D_z = 2 \times 1\{V \le 0.33\} + 1\{0.33 < V \le 0.66\}$ for $z = 0, 1, 2$, $D = \sum_{z=0}^{2} 1\{Z = z\} \times D_z$, and $Y = \sum_{d=0}^{2} 1\{D = d\} \times N_d$. All the variables U, V, N_0, N_1, and N_2 are mutually independent. Clearly, Assumption 2.2 holds in this case with $z_1 = 0$, $z_2 = 1$, and $z_3 = 2$.

Table 1 shows the results of the simulations. The rejection rates were influenced by the values of τ_n and ξ. For each measure ν, a smaller τ_n yields greater rejection rates, because a smaller τ_n leads to a smaller critical value according to (27). For $\tau_n = 2$, all the rejection rates were close to those for $\tau_n = \infty$ (the conservative case). Similar to the pattern of the results shown in Kitagawa (2015), some rejection rates for $\tau_n = 2$ with δ_ξ centered at particular values of ξ were slightly upwardly biased compared to the nominal size. Overall, however, the results showed good performance of the test in terms of size control. When sample sizes are less than or close to 3000, we suggest using $\tau_n = 2$ in practice to achieve good size control without a significant power loss. When the sample size increases, τ_n should be increased accordingly. It is also worth noting that when we used the measure $\bar{\nu}_\xi$, the rejection rates could be well controlled by the nominal significance level. Thus if we have no additional information about the choice of ξ, $\bar{\nu}_\xi$ can be a default choice for us.

4.2 Rejection Rates against Fixed Alternatives

The second set of simulations was designed to investigate the power of the test. Six data generating processes (DGPs) in total were considered, and Assumption 2.2 did not hold with $z_1 = 0$, $z_2 = 1$, and $z_3 = 2$. Sample sizes were set to $n = 200$, 600, 1000, 1100, and 2000. The

Table 1: Rejection Rates under H_0 for Multivalued D and Multivalued Z

τ_n	ξ for δ_ξ										$\bar{\nu}_\xi$
	0.07	0.1	0.13	0.16	0.19	0.22	0.25	0.28	0.3	1	
0.1	0.122	0.108	0.096	0.096	0.108	0.092	0.092	0.092	0.092	0.092	0.108
0.5	0.092	0.070	0.068	0.074	0.064	0.069	0.069	0.069	0.069	0.069	0.075
1	0.079	0.060	0.047	0.068	0.056	0.058	0.061	0.061	0.061	0.061	0.054
2	0.073	0.050	0.037	0.050	0.050	0.055	0.048	0.048	0.048	0.048	0.047
3	0.073	0.048	0.037	0.050	0.050	0.049	0.048	0.048	0.048	0.048	0.047
4	0.073	0.048	0.037	0.050	0.050	0.049	0.048	0.048	0.048	0.048	0.047
∞	0.073	0.048	0.037	0.050	0.050	0.049	0.048	0.048	0.048	0.048	0.047

probability $\mathbb{P}(Z = 2) = r_n$, with $r_n = 1/2$, $1/6$, $1/2$, $1/11$, and $1/2$ for the corresponding sample sizes. We set τ_n to 2, as suggested in the preceding set of simulations. DGPs (1)–(4) are the cases where (3) was violated and (4) was not violated, and DGPs (5) and (6) are the cases where both (3) and (4) were violated. We let $U \sim \mathrm{Unif}(0,1)$, $V \sim \mathrm{Unif}(0,1)$, $W \sim \mathrm{Unif}(0,1)$, and $Z = 2 \times 1\{U \le r_n\} + 1\{r_n < U \le r_n + 0.2\}$.

For DGPs (1)–(4), we let $D_z = 2 \times 1\{V \le 0.45\} + 1\{0.45 < V \le 0.55\}$ for $z = 0, 1, 2$, $D = \sum_{z=0}^{2} 1\{Z = z\} \times D_z$, $N_{00} \sim \mathrm{N}(0,1)$, $N_{10} \sim \mathrm{N}(0,1)$, and $N_{dz} \sim \mathrm{N}(0,1)$ for $d = 0, 1, 2$ and $z = 1, 2$.

(1): $N_{20} \sim \mathrm{N}(-0.7, 1)$ and $Y = \sum_{z=0}^{2} 1\{Z = z\} \times (\sum_{d=0}^{2} 1\{D = d\} \times N_{dz})$.

(2): $N_{20} \sim \mathrm{N}(0, 1.675^2)$ and $Y = \sum_{z=0}^{2} 1\{Z = z\} \times (\sum_{d=0}^{2} 1\{D = d\} \times N_{dz})$.

(3): $N_{20} \sim \mathrm{N}(0, 0.515^2)$ and $Y = \sum_{z=0}^{2} 1\{Z = z\} \times (\sum_{d=0}^{2} 1\{D = d\} \times N_{dz})$.

(4): $N_{20a} \sim \mathrm{N}(-1, 0.125^2)$, $N_{20b} \sim \mathrm{N}(-0.5, 0.125^2)$, $N_{20c} \sim \mathrm{N}(0, 0.125^2)$, $N_{20d} \sim \mathrm{N}(0.5, 0.125^2)$, $N_{20e} \sim \mathrm{N}(1, 0.125^2)$, $N_{20} = 1\{W \le 0.15\} \times N_{20a} + 1\{0.15 < W \le 0.35\} \times N_{20b} + 1\{0.35 < W \le 0.65\} \times N_{20c} + 1\{0.65 < W \le 0.85\} \times N_{20d} + 1\{W > 0.85\} \times N_{20e}$, and $Y = \sum_{z=0}^{2} 1\{Z = z\} \times (\sum_{d=0}^{2} 1\{D = d\} \times N_{dz})$.

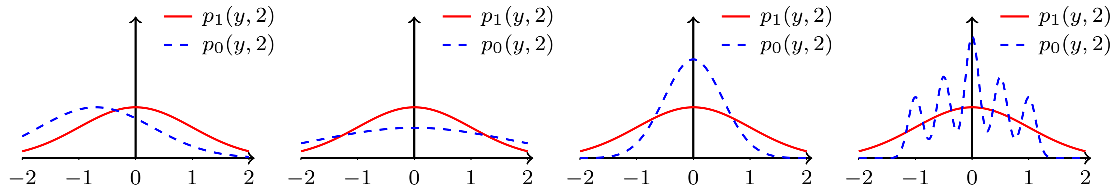
For DGPs (5) and (6), we let $N_0 \sim \mathrm{N}(0,1)$, $N_1 \sim \mathrm{N}(1,1)$, and $N_2 \sim \mathrm{N}(2,1)$.

(5): $D_0 = 2 \times 1\{V \le 0.6\} + 1\{0.6 < V \le 0.8\}$, $D_1 = 2 \times 1\{V \le 0.33\} + 1\{0.33 < V \le 0.66\}$, $D_2 = D_1$, $D = \sum_{z=0}^{2} 1\{Z = z\} \times D_z$, and $Y = \sum_{d=0}^{2} 1\{D = d\} \times N_d$.

(6): $D_0 = 2 \times 1\{V \le 0.33\} + 1\{0.33 < V \le 0.66\}$, $D_1 = 2 \times 1\{V \le 0.6\} + 1\{0.6 < V \le 0.8\}$, $D_2 = D_0$, $D = \sum_{z=0}^{2} 1\{Z = z\} \times D_z$, and $Y = \sum_{d=0}^{2} 1\{D = d\} \times N_d$.

All the variables U, V, N_{00}, N_{10}, N_{20}, N_{01}, N_{11}, N_{21}, N_{02}, N_{12}, N_{22}, N_0, N_1, and N_2 were set to be mutually independent. We briefly explain how DGPs (1)–(4) violate (3), which is shown graphically in Figure 2. We let $p_z(y, d)$ be the derivative of $\mathbb{P}(Y \in (-\infty, y], D = d | Z = z)$ with respect to y for all $d, z \in \{0, 1, 2\}$. Similar to Figure 1, if (3) were true, then we would have $p_0(y,2) \leq p_1(y,2) \leq p_2(y,2)$ everywhere. For DGPs (1)–(4), $p_1(y,2) = p_2(y,2)$ held for all y, but $p_0(y,2) \leq p_1(y,2)$ did not hold on some range of $\mathbb{R}$. DGPs (5) and (6) are the cases where the monotonicity assumption did not hold and both (3) and (4) were violated.

Figure 2: Curves of $p_0(y,2)$ (dashed) and $p_1(y,2)$ (solid) for DGPs (1)–(4)



(a) DGP (1) (b) DGP (2) (c) DGP (3) (d) DGP (4)

Table 2 shows the rejection rates under DGPs (1)–(6), that is, the power of the test. For each DGP and each measure ν, the rejection rate increased as the sample size n was increased. The results for $\nu = \bar{\nu}_\xi$ showed that if we have no information about the choice of ξ, using the weighted average of the statistics over ξ is a desirable option. When $n > 200$, the rejection rates for using $\nu = \bar{\nu}_\xi$ were at a relatively high level compared to the results for using a Dirac measure.

4.3 Nonsharpness of the Proposed Testable Implications

As discussed above, when the treatment or the instrument is multivalued, no evidence has been found that the proposed testable implications are sharp for the IV validity assumptions. In this section, we numerically examine how close the testable implications proposed in the present paper are to the sharp ones. We randomly drew DGPs from Dirichlet process distributions and assessed the proportions of the DGPs for which the proposed testable implications do not refute the IV validity conditions but the linear programming based testable implications of Balke and Pearl (1997) refute.

Specifically, let $Y \in \mathcal{Y} = \{1, \ldots, 100\}$, $D \in \mathcal{D} = \{0, 1, 2\}$, and $Z \in \mathcal{Z} = \{0, 1, 2, 3, 4\}$. Following the notation of Ghosal and Van der Vaart (2017), we let α be a finite positive Borel measure on $\mathbb{R}^3$. Let $|\alpha| = \alpha(\mathbb{R}^3)$, and $\bar{\alpha} = \alpha/|\alpha|$ which is a probability measure obtained by normalizing α. We write $P \sim \mathrm{DP}(\alpha)$ to indicate that P has a Dirichlet process distribution with base measure α.

We randomly generated 10 base measures α such that $\bar{\alpha}$ is a discrete probability measure on $\mathcal{Y} \times \mathcal{D} \times \mathcal{Z}$. For each α, we generated 1000 probability distributions from the Dirichlet process distribution $\mathrm{DP}(\alpha)$ following the procedure in Section 4.2.4 of Ghosal and Van der

Table 2: Rejection Rates under H_1 for Multivalued D and Multivalued Z

DGP	n	ξ for δ_ξ										$\bar{\nu}_\xi$
		0.07	0.1	0.13	0.16	0.19	0.22	0.25	0.28	0.3	1	
(1)	200	0.060	0.140	0.175	0.200	0.185	0.155	0.153	0.153	0.153	0.153	0.159
	600	0.672	0.683	0.616	0.482	0.323	0.230	0.214	0.214	0.214	0.214	0.516
	1000	0.606	0.729	0.790	0.792	0.775	0.738	0.715	0.715	0.715	0.715	0.777
	1100	0.889	0.859	0.720	0.504	0.314	0.216	0.217	0.217	0.217	0.217	0.658
	2000	0.969	0.988	0.993	0.987	0.989	0.979	0.975	0.975	0.975	0.975	0.991
(2)	200	0.030	0.060	0.074	0.076	0.076	0.069	0.072	0.072	0.072	0.072	0.064
	600	0.347	0.168	0.069	0.054	0.059	0.059	0.056	0.056	0.056	0.056	0.083
	1000	0.404	0.379	0.294	0.146	0.088	0.059	0.062	0.062	0.062	0.062	0.153
	1100	0.434	0.123	0.054	0.059	0.059	0.059	0.060	0.060	0.060	0.060	0.084
	2000	0.896	0.897	0.775	0.521	0.269	0.177	0.154	0.154	0.154	0.154	0.635
(3)	200	0.087	0.177	0.240	0.307	0.325	0.297	0.290	0.290	0.290	0.290	0.262
	600	0.695	0.719	0.728	0.693	0.577	0.466	0.434	0.434	0.434	0.434	0.673
	1000	0.660	0.743	0.826	0.856	0.880	0.887	0.875	0.875	0.875	0.875	0.878
	1100	0.884	0.924	0.899	0.773	0.622	0.516	0.517	0.517	0.517	0.517	0.840
	2000	0.968	0.985	0.991	0.995	0.995	0.998	0.999	0.999	0.999	0.999	0.999
(4)	200	0.038	0.099	0.147	0.155	0.148	0.138	0.135	0.135	0.135	0.135	0.146
	600	0.402	0.376	0.366	0.290	0.207	0.209	0.189	0.189	0.189	0.189	0.304
	1000	0.331	0.433	0.407	0.406	0.444	0.475	0.477	0.477	0.477	0.477	0.483
	1100	0.498	0.526	0.492	0.355	0.203	0.137	0.137	0.137	0.137	0.137	0.403
	2000	0.597	0.704	0.710	0.725	0.741	0.769	0.791	0.791	0.791	0.791	0.796
(5)	200	0.365	0.487	0.589	0.626	0.685	0.752	0.780	0.780	0.780	0.780	0.699
	600	0.980	0.990	0.995	0.997	0.998	0.998	0.998	0.998	0.998	0.998	0.998
	1000	0.994	0.998	0.999	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	1100	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	2000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
(6)	200	0.372	0.482	0.545	0.616	0.659	0.701	0.711	0.711	0.711	0.711	0.664
	600	0.704	0.823	0.904	0.929	0.962	0.981	0.988	0.988	0.988	0.988	0.965
	1000	0.992	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	1100	0.912	0.957	0.979	0.984	0.990	0.995	0.995	0.995	0.995	0.995	0.990
	2000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Vaart (2017). Specifically, for every measure α, let $(Y_1, D_1, Z_1) \sim \bar{\alpha}$ and

$$(Y_{m+1}, D_{m+1}, Z_{m+1})|(Y_1, D_1, Z_1), \ldots, (Y_m, D_m, Z_m) \sim \frac{\alpha + \sum_{i=1}^{m} \delta_{(Y_i, D_i, Z_i)}}{|\alpha| + m}$$

$$\text{for all } 1 \leq m \leq n-1, \tag{34}$$

where $\delta_{(Y_i,D_i,Z_i)}$ denotes the Dirac measure centered at (Y_i, D_i, Z_i) and $n = 1000$. We repeated this procedure 1000 times and obtained 1000 empirical distributions. We then verified whether the generated empirical distributions satisfied the linear programming based bound conditions of Balke and Pearl (1997) and the proposed testable implications in this paper. Next we calculated the proportions of the distributions that were refuted by the proposed testable implications compared to all the distributions that were refuted by the bound conditions of Balke and Pearl (1997). Table 3 presents the total number (TR) of random distributions refuted by the bound conditions of Balke and Pearl (1997), and the ratio (RR) of distributions refuted by the proposed testable implications out of the TR distributions. The simulation results show that all the randomly generated distributions refuted by the bound conditions of Balke and Pearl (1997) are also refuted by the proposed testable implications. These results provide evidence that the proposed testable implications are close to the sharp ones, at least in this design.

Table 3: Nonsharpness of the Proposed Testable Implications

α	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
TR	1000	964	971	1000	959	989	999	994	999	982
RR	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

5 Empirical Application

We revisit one empirical example discussed by Kitagawa (2015) to show the performance of the proposed test in practice. The example is from Card (1993), who used college proximity as an instrument of years of schooling to study the causal link between education and earnings. The data are from the Young Men Cohort of the National Longitudinal Survey. In the original study of Card (1993), the educational level D is a multivalued treatment variable, while Kitagawa (2015) treated it as a binary treatment variable T with $T = 1\{D \geq 16\}$. The results of the test of Kitagawa (2015) showed that the instrument was not valid when no covariates were controlled.

We use the originally defined treatment variable D to reconduct the test. Specifically, the treatment D is education attainment observed in 1976 (the variable "ed76"), the instrument Z is whether an individual grew up near a 4-year college (the variable "nearc4"), and the outcome is log wage observed in 1976 (the variable "lwage76") in the data set. The available sample size is 3010. We follow the setup in Section 3 with $\mathcal{D} = \{1, 2, \ldots, 18\}$ and $\mathcal{Z} = \{0, 1\}$. The instrument $Z = 1$ implies that an individual grew up near a 4-year college. Table 4 shows the p-values obtained from our test using each measure ν. From these results, we conclude that we do not reject the validity of instrument Z. In Section E.5, we show more results by using application-based simulations to choose Ξ and ν. The results are similar to those in Table 4.

The testable implication used by Kitagawa (2015) for binary T is that

$$\begin{aligned} &\mathbb{P}(Y \in B, T = 0 | Z = 1) - \mathbb{P}(Y \in B, T = 0 | Z = 0) \leq 0 \\ &\text{and } \mathbb{P}(Y \in B, T = 1 | Z = 1) - \mathbb{P}(Y \in B, T = 1 | Z = 0) \geq 0 \end{aligned} \tag{35}$$

Table 4: p-values Obtained from the Proposed Test for Each Measure ν

ξ for δ_ξ										$\bar{\nu}_\xi$
0.07	0.1	0.13	0.16	0.19	0.22	0.25	0.28	0.3	1	
0.958	0.975	0.975	0.975	0.975	0.975	0.975	0.975	0.975	0.975	0.973

for all closed intervals B. The inequalities in (35) are equivalent to the following for all closed intervals B:

$$\begin{aligned}&\mathbb{P}(Y\in B, D<16|Z=1)-\mathbb{P}(Y\in B, D<16|Z=0)\leq 0\\ &\text{and } \mathbb{P}(Y\in B, D\geq 16|Z=1)-\mathbb{P}(Y\in B, D\geq 16|Z=0)\geq 0,\end{aligned} \tag{36}$$

which are different from those in the testable implication given in (3) and (4) and are not implied by Assumption 2.2. Thus a valid instrument Z for multivalued D which satisfies the testable implication given in (3) and (4) may not satisfy the inequalities in (35), that is, Z may not remain valid for binary (or coarsened) T. This provides a possible explanation for why we accept Z but Kitagawa (2015) rejected it.

To be more explicit, we consider a simpler example. Let $U\sim \text{Unif}(0,1)$, $V\sim\text{Unif}(0,1)$, $Y_d\sim\text{Unif}(d,d+1)$ for $d\in\{0,1,2\}$, $Z=1\{U\leq 0.5\}$, $D=\sum_{z=0}^{1}1\{Z=z\}\times D_z$ with $D_0=2\times 1\{V\leq 0.1\}+1\{0.1<V\leq 0.5\}$ and $D_1=2\times 1\{V\leq 0.5\}+1\{0.5<V\leq 0.6\}$, and $Y=\sum_{d=0}^{2}1\{D=d\}\times Y_d$, where U, V, Y_0, Y_1, and Y_2 are mutually independent. We can verify that Assumption 2.2 holds for Z and D in this example. It can be shown that for every Borel set B and each $z\in\{0,1\}$, $\mathbb{P}(Y\in B, D\geq 1|Z=z)=\mathbb{P}(Y_1\in B, D_z=1)+\mathbb{P}(Y_2\in B, D_z=2)$. Let $B=[1,2]$. Then we have

$$\begin{aligned}&\mathbb{P}(Y\in B, D\geq 1|Z=1)-\mathbb{P}(Y\in B, D\geq 1|Z=0)\\ =&\mathbb{P}(D_0=0, D_1=1)-\mathbb{P}(D_0=1, D_1=2)<0.\end{aligned} \tag{37}$$

The inequality in (37) shows that the valid instrument Z for multivalued D does not satisfy the inequalities as those in (36). Equivalently, the instrument Z is not valid for the coarsened treatment $T=1\{D\geq 1\}$. The reason why Z does not remain valid is as follows. Assumption 2.1 for Z and T specified in this example requires $Y'_{t0}=Y'_{t1}$ almost surely for $t\in\{0,1\}$, where Y'_{tz} is the potential outcome variable for $T=t$ and $Z=z$ with $t\in\{0,1\}$ and $z\in\{0,1\}$. With the potential outcome variables, we can write

$$Y=\sum_{d=0}^{2}1\{D=d\}\cdot Y_d=\sum_{z=0}^{1}1\{Z=z\}\cdot\left(\sum_{t=0}^{1}1\{T=t\}\cdot Y'_{tz}\right).$$

For every $\omega\in\Omega$ with $Z(\omega)=z$ and $T(\omega)=1$, $Y(\omega)=\sum_{d=1}^{2}1\{D_z(\omega)=d\}\cdot Y_d(\omega)=$

$Y'_{1z}(\omega)$. If $Y'_{10} = Y'_{11}$ almost surely, it follows that

$$Y'_{10} = Y'_{11} = \sum_{d=1}^{2} 1\{D = d\} \cdot Y_d + 1\{D = 0\} \cdot W \text{ almost surely with } D = \sum_{z=0}^{1} 1\{Z = z\} \cdot D_z, \tag{38}$$

where W is a random variable such that $W(\omega) = Y'_{10}(\omega) = Y'_{11}(\omega)$ for almost all ω with $T(\omega) = 0$. However, (38) shows that Z affects Y'_{10} and Y'_{11} through D, and therefore Y'_{10} and Y'_{11} are not necessarily independent of Z. Thus Assumption 2.1(ii) may fail for Z and (coarsened) T. This verifies the rejection conclusion of Kitagawa (2015).

For empirical or theoretical reasons, we may want to coarsen a multivalued treatment to be a binary variable in some circumstances. However, Angrist and Imbens (1995, p. 436) and Marshall (2016) showed that such coarsening may lead to inconsistent estimates for the average per-unit treatment effect and the effect of obtaining a particular treatment intensity level beyond obtaining only the preceding level. They provided several special cases in which the estimates could be consistent, such as the case where the instrument only affects reaching a particular treatment intensity and the case where the effect at all intensities other than a particular one is zero. But further discussion of Marshall (2016) showed that these cases are often implausible in practice. For the data set of Card (1993), the coarsened treatment variable, $T = 1\{D \geq 16\}$, can be considered as a four-year college degree. The simple numerical example designed above shows that coarsening may undermine the validity of the instrument for T, so the IV estimate for the effect of obtaining a college degree may be inconsistent. This provides another perspective for understanding the inconsistency of the coarsened estimator, and shows the significance of the generalization of the test in the present paper.

In the empirical example of Card (1993), the numbers of observations for $D = d_{\min}$ and $D = d_{\max}$ are small, so we mainly rely on (4) to examine the IV validity assumption. This may be another reason why the proposed test does not reject the null hypothesis. In practice, when the number of observations for $D = d_{\min}$ or $D = d_{\max}$ is small, we may redefine the treatment as Kitagawa (2015) did ($T = 1\{D \geq 16\}$). In this case, we need to make sure that the treatment used in the empirical analysis and the treatment used in the test are the same, because the definitions of all types of individuals (always takers, compliers, defiers, and never takers) would be different if we define the treatment in different ways. In applications, we suggest that users first clarify the definition of the treatment, and then choose a test for the validity of the instrument appropriately. For a multivalued treatment in interest, we suggest using the proposed test without coarsening the treatment. If the coarsened version of the treatment is of interest instead, we should adjust the testing approach accordingly.

5.1 Further Discussion on IV Validity Assumptions

Assumption 2.2 allows for multiple instruments, that is, Z can be a vector instrument. Theorems 1 and 2 of Angrist and Imbens (1995) show that under Assumption 2.2, the two-stage least square (2SLS) estimand can be interpreted as a weighted average of average causal responses, whether Z is multiple or not. Mogstad et al. (2021) show that when Z is multiple (a vector instrument), the monotonicity condition in Assumption 2.2 holds only when individual choice behavior is effectively homogeneous. Such a homogeneity assumption, however, may be restrictive in applications. Mogstad et al. (2021) consider the example where the treatment is college attendance and the instruments are tuition and proximity to college. Specifically, suppose that $Z = (Z_1, Z_2)$, where Z_1 is a tuition subsidy and Z_2 is proximity to a college, and that $Z \in \{(0,0),(0,1),(1,0),(1,1)\}$. The monotonicity condition in Assumption 2.2 requires that all individuals respond more to tuition than to proximity, or vice versa. That is, $D_{(0,1)} \leq D_{(1,0)}$ almost surely, or $D_{(1,0)} \leq D_{(0,1)}$ almost surely. This requirement, however, may be strong in practice, since some individuals may prefer tuition subsidy and other individuals may prefer proximity to a college. If such homogeneity does not hold, then Assumption 2.2 is violated. As a consequence, the identification and the estimation results of causal effects may not hold by Theorems 1 and 2 of Angrist and Imbens (1995). That is, the average causal responses may not be identified, and the 2SLS estimand may not be interpreted as a weighted average of causal effects.

To salvage the results under the restrictive monotonicity condition in Assumption 2.2, Mogstad et al. (2021) provide a weaker version of monotonicity condition, partial monotonicity (Assumption PM in Mogstad et al. (2021)). Partial monotonicity is a condition that only compares values of a single component of the instrument, fixing all other components. Consider $\mathcal{Z} \subset \mathbb{R}^L$. Mogstad et al. (2021) divide $z \in \mathcal{Z}$ into its lth component z_l and all other $(L-1)$ components z_{-l}. Then z can be written as $z = (z_l, z_{-l})$. The partial monotonicity requires that for every $l \in \{1,\ldots,L\}$, and all (z_l, z_{-l}) and (z_l', z_{-l}), either $D_{(z_l,z_{-l})} \geq D_{(z_l',z_{-l})}$ almost surely, or $D_{(z_l,z_{-l})} \leq D_{(z_l',z_{-l})}$ almost surely. Under this weaker monotonicity condition and other relevant conditions, Mogstad et al. (2021) show that the 2SLS estimand can still have a causal interpretation. Mogstad et al. (2020) develop a new marginal treatment effect framework under the partial monotonicity condition, and based on this framework they develop a new methodology for partial identification of policy-relevant target parameters.[27]

Lee and Salanié (2018) provide identifying conditions for treatment effects in a more

[27] As discussed in Carr and Kitagawa (2021), their testing approach can be extended to the partial monotonicity of Mogstad et al. (2021).

general class of models, where the usual monotonicity assumption is relaxed. The treatment is allowed to be ordered or unordered in their models. As discussed in Section 5.3 of Lee and Salanié (2018), the unordered monotonicity assumption, Assumption 2.3 (Assumption A-3 of Heckman and Pinto (2018)), can be rephrased in the notation of Lee and Salanié (2018), and obeys their identifying assumptions.[28] As stated in Heckman and Pinto (2018), the unordered monotonicity condition requires that as the instrument value shifts from one to another, all agents move uniformly toward or against each possible treatment value. This requirement may not be satisfied in Example 1 (Selection with Two-way Flows) in Lee and Salanié (2018). Lee and Salanié (2018) (Example 1) consider a two-part exam: Failing both parts assigns individuals to the treatment $D = 0$, passing both to $D = 1$, and failing only one to $D = 2$. A shift in the value of the instrument Z may increase both the minimum grades of the two-part exam. As illustrated in Figure 1 of Lee and Salanié (2018), individuals move not uniformly toward or against each value of D because of the two dimensions of the criteria. Specifically, let Z be a discrete random variable. Suppose that there are two random variables V_1 and V_2 and two threshold functions Q_1 and Q_2 such that for every $z \in \mathcal{Z}$,

$$\begin{aligned}&D_z = 0 \text{ if } V_1 < Q_1(z) \text{ and } V_2 < Q_2(z), D_z = 1 \text{ if } V_1 \geq Q_1(z) \text{ and } V_2 \geq Q_2(z),\\&\text{and } D_z = 2 \text{ otherwise.}\end{aligned}$$

Let $D = \sum_z 1\{Z = z\}D_z$. Suppose the probability space is $(\Omega, \mathcal{A}, \mathbb{P})$. Let the value of Z change from z to z' such that $Q_j(z) < Q_j(z')$ for $j \in \{1, 2\}$. If $D_z(\omega) = 2$ $(V_1(\omega) \geq Q_1(z)$ and $V_2(\omega) < Q_2(z))$, then it is possible that $D_{z'}(\omega) = 0$ $(V_1(\omega) < Q_1(z')$ and $V_2(\omega) < Q_2(z'))$ or that $D_{z'}(\omega) = 2$ $(V_1(\omega) \geq Q_1(z')$ and $V_2(\omega) < Q_2(z'))$. If $D_{z'}(\omega) = 2$ $(V_1(\omega) \geq Q_1(z')$ and $V_2(\omega) < Q_2(z'))$, then it is possible that $D_z(\omega) = 1$ $(V_1(\omega) \geq Q_1(z)$ and $V_2(\omega) \geq Q_2(z))$ or that $D_z(\omega) = 2$ $(V_1(\omega) \geq Q_1(z)$ and $V_2(\omega) < Q_2(z))$. Therefore, the unordered monotonicity assumption may fail in this example. Consequently, the identifying results of Heckman and Pinto (2018) (Theorem T-6) which rely on the unordered monotonicity may not hold. The selection with two-way flows example can be incorporated into the framework of Lee and Salanié (2018).[29] Therefore, Lee and Salanié (2018) allow for more general treatment effect models.

As discussed above, the IV validity assumptions are restrictive and may be violated in different settings, and canonical identification and estimation results may break down as a consequence. The proposed testing method helps falsify these conditions and provide evidence for the plausibility of relevant causal effect analysis. If the IV validity assumptions

[28] The only difference is that Lee and Salanié (2018) require continuous instruments, but Heckman and Pinto (2018) do not.

[29] See also Example 2 (Entry Game) in Lee and Salanié (2018).

are refuted, we can move forward and apply other novel approaches, such as those in Lee and Salanié (2018), Mogstad et al. (2020), and Mogstad et al. (2021).

6 Conclusion

In this paper, we provided a general framework for testing instrument validity in heterogeneous causal effect models. We generalized the testable implications of the instrument validity assumptions in the literature, and based on them we proposed a nonparametric bootstrap test. An extended continuous mapping theorem and an extended delta method were provided to establish the asymptotic distribution of the test statistic, which may be of independent interest. The proposed test can be applied in more general settings and may achieve power improvement.

Appendix

A Extended Continuous Mapping Theorem and Extended Delta Method

We follow van der Vaart and Wellner (1996) to introduce some notation we use multiple times in the appendix. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be an arbitrary probability space. For an arbitrary map $T : \Omega \to \bar{\mathbb{R}}$, we define the outer integral or outer expectation of T with respect to $\mathbb{P}$ by

$$E^* [T] = \inf \left\{ E[U] : U \geq T, U : \Omega \to \bar{\mathbb{R}} \text{ measurable and } E[U] \text{ exists} \right\}.$$

The outer probability of an arbitrary subset B of Ω is

$$\mathbb{P}^* (B) = \inf \{ \mathbb{P}(A) : A \supset B, A \in \mathcal{A} \}.$$

The inner integral (or inner expectation) and the inner probability can be defined as

$$E_* [T] = -E^* [-T] \text{ and } \mathbb{P}_* (B) = 1 - \mathbb{P}^* (\Omega \setminus B),$$

respectively. We denote a minimal measurable majorant of T (resp. a maximal measurable minorant) by T^* (resp. T_*), which always exists by Lemma 1.2.1 of van der Vaart and Wellner (1996). Suppose T is a real-valued map defined on an arbitrary product probability space $(\Omega_1 \times \Omega_2, \mathcal{A}_1 \times \mathcal{A}_2, \mathbb{P}_1 \times \mathbb{P}_2)$. We write $E^*[T]$ for the outer expectation as before, and for every ω_1, we define

$$E_2^*[T](\omega_1) = \inf \int U(\omega_2)\, \mathrm{d}\mathbb{P}_2(\omega_2), \tag{A.1}$$

where the infimum is taken over all measurable functions $U : \Omega_2 \to \bar{\mathbb{R}}$ with $U(\omega_2) \geq T(\omega_1, \omega_2)$ for all ω_2 such that $\int U\, \mathrm{d}\mathbb{P}_2$ exists. Then $E_1^*[E_2^*[T]]$ is the outer integral of the function $E_2^*[T] : \Omega_1 \to \bar{\mathbb{R}}$, and we call $E_1^*[E_2^*[T]]$ the repeated outer expectation. We define the repeated inner expectation $E_{1*}[E_{2*}[T]]$ analogously.[30]

Theorem A.1 (Extended continuous mapping) *Let $\mathbb{D}$ and $\mathbb{E}$ be metric spaces with metrics d and e, respectively. Let $\mathbb{D}_0 \subset \mathbb{D}$. Let X be Borel measurable and take values in $\mathbb{D}_0$. Suppose, in addition, that either of the following conditions holds:*

(a) Let $\mathbb{D}_n \subset \mathbb{D}$. Let $X_n : \Omega \to \mathbb{D}$ with $X_n(\omega) \in \mathbb{D}_n$ for all $\omega \in \Omega$ and all n. Let g_n be a random map with $g_n(\omega) : \mathbb{D}_n \to \mathbb{E}$ (for every $\omega \in \Omega$, $g_n(\omega)$ is a map on $\mathbb{D}_n$). The random map g_n satisfies the condition that for every $\varepsilon > 0$ there is a measurable set $A \subset \Omega$ with $\mathbb{P}(A) \geq 1 - \varepsilon$ such that if $x_n \to x$ with $x_n \in \mathbb{D}_n$ and $x \in \mathbb{D}_0$, then $g_n(x_n)$ converges to $g(x)$ uniformly on A ($\sup_{\omega \in A} e(g_n(\omega)(x_n), g(x)) \to 0$),[31] where $g : \mathbb{D}_0 \to \mathbb{E}$ is a fixed (deterministic) map. Also, X is separable.

(b) Let $\mathbb{D}_n(\omega) \subset \mathbb{D}$ for all $\omega \in \Omega$ and all n. Let $X_n : \Omega \to \mathbb{D}$ with $X_n(\omega) \in \mathbb{D}_n(\omega)$ for all $\omega \in \Omega$ and all n. Let g_n be a random map with $g_n(\omega) : \mathbb{D}_n(\omega) \to \mathbb{E}$ (for every $\omega \in \Omega$, $g_n(\omega)$ is a map on $\mathbb{D}_n(\omega)$). The random map g_n satisfies the condition that for every $\varepsilon > 0$ there is a measurable set $A \subset \Omega$ with $\mathbb{P}(A) \geq 1 - \varepsilon$ such that for every subsequence $\{x_{n_m}\}$, if $x_{n_m} \to x$ with $x_{n_m} \in \mathbb{D}_{n_m}(\omega_{n_m})$, $\omega_{n_m} \in A$, and $x \in \mathbb{D}_0$, then $g_{n_m}(\omega_{n_m})(x_{n_m})$ converges to $g(x)$, where $g : \mathbb{D}_0 \to \mathbb{E}$ is a fixed continuous map.

Then we have that

(i) $X_n \rightsquigarrow X$ implies that $g_n(X_n) \rightsquigarrow g(X)$;

(ii) If X_n converges to X in outer probability,[32] then $g_n(X_n)$ converges to $g(X)$ in outer probability;

(iii) If X_n converges to X outer almost surely,[33] then $g_n(X_n)$ converges to $g(X)$ outer almost surely.

[30] Additional technical details about the repeated expectations can be found in van der Vaart and Wellner (1996, pp. 10–12).

[31] This is a condition similar to almost uniform convergence. See Definition 1.9.1(ii) of van der Vaart and Wellner (1996). By Lemma 1.9.2(iii) of van der Vaart and Wellner (1996), almost uniform convergence is equivalent to outer almost sure convergence if the limit is Borel measurable.

[32] See Definition 1.9.1(i) of convergence in outer probability in van der Vaart and Wellner (1996).

[33] See Definition 1.9.1(iii) of outer almost sure convergence in van der Vaart and Wellner (1996).

Remark A.1 *Theorem A.1 is an extension of Theorem 1.11.1 (extended continuous mapping) of van der Vaart and Wellner (1996). Theorem 1.11.1 of van der Vaart and Wellner (1996) assumes that every g_n is a fixed map. Theorem A.1 allows every g_n to be random. Theorem A.1(i) will be used to establish Theorem A.2 (extended delta method).*

Proof of Theorem A.1. Suppose Condition (a) holds. Assume the weakest of the three assumptions: the one in (i) that $X_n \rightsquigarrow X$. **First**, let $\mathbb{D}_\infty$ be the set of all x for which there exists a sequence $\{x_n\}$ with $x_n \in \mathbb{D}_n$ and $x_n \to x$. By the representation theorem (see, for example, Theorem 9.4 of Pollard (1990) or Theorem 1.10.4 of van der Vaart and Wellner (1996)), along the lines of the second paragraph in the proof of Theorem 1.11.1 of van der Vaart and Wellner (1996), we can show that $\mathbb{P}_*(X \in \mathbb{D}_\infty) = 1$. **Second**, fix ε and a measurable set A with $\mathbb{P}(A) \geq 1-\varepsilon$ that satisfies the assumptions, and suppose there is some subsequence such that $x_{n'} \to x$ with $x_{n'} \in \mathbb{D}_{n'}$ for all n' and $x \in \mathbb{D}_0 \cap \mathbb{D}_\infty$. Since $x \in \mathbb{D}_\infty$, there is a sequence $y_n \to x$ with $y_n \in \mathbb{D}_n$ for all n. Fill out the subsequence $x_{n'}$ to an entire sequence by putting $x_n = y_n$ for all $n \notin \{n'\}$. Then by assumption, $g_n(x_n) \to g(x)$ uniformly on A on this entire sequence, hence also on the subsequence, that is, $g_{n'}(x_{n'}) \to g(x)$ uniformly on A. **Third**, let $x_m \to x$ in $\mathbb{D}_0 \cap \mathbb{D}_\infty$. For every m, there is a sequence $y_{m,n} \in \mathbb{D}_n$ with $y_{m,n} \to x_m$ as $n \to \infty$. Fix a small $\varepsilon > 0$ and a measurable set A with $\mathbb{P}(A) \geq 1-\varepsilon$ that satisfies the assumptions. Now we have that $g_n(y_{m,n}) \to g(x_m)$ uniformly on A. For every m, take n_m such that $|y_{m,n_m} - x_m| < 1/m$ and $|g_{n_m}(y_{m,n_m}) - g(x_m)| < 1/m$ uniformly on A and such that n_m is increasing in m. Then $y_{m,n_m} \to x$, and hence $g_{n_m}(y_{m,n_m}) \to g(x)$ uniformly on A. Since $|g(x_m) - g(x)| \leq |g_{n_m}(y_{m,n_m}) - g(x_m)| + |g_{n_m}(y_{m,n_m}) - g(x)|$ uniformly on A, we have $|g(x_m) - g(x)| \to 0$. Thus g is continuous on $\mathbb{D}_0 \cap \mathbb{D}_\infty$.

For simplicity of notation, we will write $\mathbb{D}_0$ for $\mathbb{D}_0 \cap \mathbb{D}_\infty$. Without loss of generality, we assume that X takes its values in $\mathbb{D}_0$. Since g is continuous on $\mathbb{D}_0$ now, $g(X)$ is Borel measurable.

(i). Let F be an arbitrary closed set in $\mathbb{E}$. By the assumptions, for every $\varepsilon > 0$ there is a measurable set $A \subset \Omega$ with $\mathbb{P}(A) \geq 1-\varepsilon$ such that if $x_n \to x$ with $x_n \in \mathbb{D}_n$ and $x \in \mathbb{D}_0$, then $g_n(x_n)$ converges to $g(x)$ uniformly on A, that is, $\sup_{\omega \in A} |g_n(\omega)(x_n) - g(x)| \to 0$. Fix ε and A. Then

$$\cap_{k=1}^{\infty} \overline{\cup_{m=k}^{\infty} \cup_{\omega \in A} \left(g_m(\omega)\right)^{-1}(F)} \subset g^{-1}(F) \cup (\mathbb{D} - \mathbb{D}_0). \tag{A.2}$$

Suppose x is an element of the set on the left-hand side of (A.2). For every n, there exist $n' > n$, $\omega_{n'} \in A$, and $x_{n'} \in g_{n'}(\omega_{n'})^{-1}(F) \subset \mathbb{D}_{n'}$ such that $d(x_{n'}, x) \leq 1/n$. Therefore, there is a subsequence $x_{n_m} \in g_{n_m}(\omega_{n_m})^{-1}(F) \subset \mathbb{D}_{n_m}$ with $\omega_{n_m} \in A$ such that $n_m \uparrow \infty$ and $x_{n_m} \to x$ as $m \to \infty$. By the definition of A, either $g_{n_m}(\omega_{n_m})(x_{n_m}) \to g(x)$ or $x \notin \mathbb{D}_0$. Since

F is closed, this implies that $g(x) \in F$ or $x \notin \mathbb{D}_0$. Then for every k,

$$\begin{aligned}
\limsup_{n\to\infty} \mathbb{P}^*\left(g_n\left(X_n\right) \in F\right) &\leq \limsup_{n\to\infty} \mathbb{P}^*\left(\left\{\left\{X_n \in \overline{\cup_{m=k}^{\infty} g_m^{-1}\left(F\right)}\right\} \cap A\right\} \cup A^c\right) \\
&= \limsup_{n\to\infty} E\left[\left(1\left\{\left\{X_n \in \overline{\cup_{m=k}^{\infty} g_m^{-1}\left(F\right)}\right\} \cap A\right\} \vee 1\left\{A^c\right\}\right)^*\right],
\end{aligned} \tag{A.3}$$

where the equality is from Lemmas 1.2.3(i) and 1.2.1 of van der Vaart and Wellner (1996). Then by Lemmas 1.2.2(viii), 1.2.1, and 1.2.3(i) of van der Vaart and Wellner (1996),

$$\begin{aligned}
&E\left[\left(1\left\{\left\{X_n \in \overline{\cup_{m=k}^{\infty} g_m^{-1}\left(F\right)}\right\} \cap A\right\} \vee 1\left\{A^c\right\}\right)^*\right] \\
&= E\left[\left(1\left\{\left\{X_n \in \overline{\cup_{m=k}^{\infty} g_m^{-1}\left(F\right)}\right\} \cap A\right\}\right)^* \vee \left(1\left\{A^c\right\}\right)\right] \\
&\leq \mathbb{P}^*\left(\left\{X_n \in \overline{\cup_{m=k}^{\infty} g_m^{-1}\left(F\right)}\right\} \cap A\right) + \mathbb{P}\left(A^c\right).
\end{aligned} \tag{A.4}$$

By (A.3) and (A.4), together with Theorem 1.3.4(iii) (portmanteau) of van der Vaart and Wellner (1996), we have

$$\begin{aligned}
\limsup_{n\to\infty} \mathbb{P}^*\left(g_n\left(X_n\right) \in F\right) &\leq \limsup_{n\to\infty} \mathbb{P}^*\left(\left\{X_n \in \overline{\cup_{m=k}^{\infty} g_m^{-1}\left(F\right)}\right\} \cap A\right) + \mathbb{P}\left(A^c\right) \\
&\leq \limsup_{n\to\infty} \mathbb{P}^*\left(X_n \in \overline{\cup_{m=k}^{\infty} \cup_{\omega\in A} \left(g_m\left(\omega\right)\right)^{-1}\left(F\right)}\right) + \varepsilon \\
&\leq \mathbb{P}\left(X \in \overline{\cup_{m=k}^{\infty} \cup_{\omega\in A} \left(g_m\left(\omega\right)\right)^{-1}\left(F\right)}\right) + \varepsilon.
\end{aligned}$$

Letting $k \to \infty$ together with (A.2) gives

$$\begin{aligned}
\limsup_{n\to\infty} \mathbb{P}^*\left(g_n\left(X_n\right) \in F\right) &\leq \mathbb{P}\left(X \in \cap_{k=1}^{\infty}\overline{\cup_{m=k}^{\infty} \cup_{\omega\in A} \left(g_m\left(\omega\right)\right)^{-1}\left(F\right)}\right) + \varepsilon \\
&\leq \mathbb{P}\left(g\left(X\right) \in F\right) + \varepsilon.
\end{aligned}$$

Since ε can be arbitrarily small, we can conclude that $\limsup_{n\to\infty} \mathbb{P}^*\left(g_n\left(X_n\right) \in F\right) \leq \mathbb{P}\left(g\left(X\right) \in F\right)$. By Theorem 1.3.4(iii) of van der Vaart and Wellner (1996) again, $g_n\left(X_n\right) \rightsquigarrow g\left(X\right)$.

(ii). Choose $\delta_n \downarrow 0$ with $\mathbb{P}^*\left(d\left(X_n, X\right) \geq \delta_n\right) \to 0$. Fix $\varepsilon > 0$. Let $A \subset \Omega$ be a measurable set with $\mathbb{P}\left(A\right) \geq 1 - \varepsilon$ that satisfies the assumptions. Let $B_n(\omega)$ be the set of all x such that there is a $y \in \mathbb{D}_n$ with $d\left(y, x\right) < \delta_n$ and $e\left(g_n\left(\omega\right)\left(y\right), g\left(x\right)\right) > \varepsilon$. Let $B_n = \cup_{\omega\in A} B_n(\omega)$. Suppose $x \in B_n$ for infinitely many n. Then there are sequences $\omega_{n_m} \in A$ and $x_{n_m} \in \mathbb{D}_{n_m}$ with $x_{n_m} \to x$ such that $e\left(g_{n_m}(\omega_{n_m})\left(x_{n_m}\right), g\left(x\right)\right) > \varepsilon$ for each m. This implies that $x_{n_m} \to x$ with $x_{n_m} \in \mathbb{D}_{n_m}$ but that $g_{n_m}(x_{n_m})$ does not converge to $g(x)$ uniformly on A. Thus by assumption, $x \notin \mathbb{D}_0$. Note that $x \in \limsup B_n$ is equivalent to $x \in B_n$ for infinitely many n. Thus we can conclude that $\limsup B_n \cap \mathbb{D}_0 = \varnothing$. Since g is continuous on $\mathbb{D}_0$, $B_n \cap \mathbb{D}_0$ is relatively open in $\mathbb{D}_0$ and hence relatively Borel. This is because if $z \in \mathbb{D}_0$ is close enough to $x \in B_n \cap \mathbb{D}_0$, then $d(y,z) \leq d(y,x) + d(x,z) < \delta_n$ and $e\left(g_n\left(\omega\right)\left(y\right), g\left(z\right)\right) \geq e\left(g_n\left(\omega\right)\left(y\right), g\left(x\right)\right) - e\left(g\left(z\right), g\left(x\right)\right) > \varepsilon$. Since X takes values in $\mathbb{D}_0$ by assumption, by Lemma 1.2.3(i) of van der Vaart and Wellner (1996),

$$\mathbb{P}^*\left(X \in B_n\right) = E^*\left[1\left\{X \in B_n\right\}\right] = E\left[1\left\{X \in B_n \cap \mathbb{D}_0\right\}\right].$$

Also, by the dominated convergence theorem,

$$\begin{aligned} &E\left[1\left\{X \in B_n \cap \mathbb{D}_0\right\}\right] \leq E\left[1\left\{X \in \cup_{m=n}^{\infty}(B_m \cap \mathbb{D}_0)\right\}\right] \\ &\rightarrow E\left[1\left\{X \in \cap_{n=1}^{\infty} \cup_{m=n}^{\infty}\left(B_m \cap \mathbb{D}_0\right)\right\}\right] = \mathbb{P}\left(X \in \limsup B_n \cap \mathbb{D}_0\right) = 0. \end{aligned}$$

This implies that $\mathbb{P}^*\left(X \in B_n\right) \rightarrow 0$ as $n \rightarrow \infty$. Now we have that

$$\begin{aligned} \mathbb{P}^*\left(e(g_n(X_n), g(X)) > \varepsilon\right) \leq &\mathbb{P}^*\left(\left\{e(g_n(X_n), g(X)) > \varepsilon\right\} \cap A\right) + \mathbb{P}\left(A^c\right) \\ \leq &\mathbb{P}^*\left(X \in B_n \text{ or } d(X_n, X) \geq \delta_n\right) + \varepsilon \rightarrow \varepsilon. \end{aligned}$$

Since ε is arbitrary, the claim holds.

(iii). By Lemmas 1.9.3(i) and 1.9.2(iii) of van der Vaart and Wellner (1996), it suffices to prove that $\sup_{m \geq n} e\left(g_m\left(X_m\right), g\left(X\right)\right)$ converges to 0 in outer probability. Choose $\delta_n \downarrow 0$ with $\mathbb{P}^*\left(\sup_{m \geq n} d\left(X_m, X\right) \geq \delta_n\right) \rightarrow 0$. Fix $\varepsilon > 0$. Let $A \subset \Omega$ be a measurable set with $\mathbb{P}\left(A\right) \geq 1 - \varepsilon$ such that if $x_n \rightarrow x$ with $x_n \in \mathbb{D}_n$ and $x \in \mathbb{D}_0$, then $g_n\left(x_n\right)$ converges to $g\left(x\right)$ uniformly on A. Let $B_n(\omega)$ be the set of all x such that there are $m \geq n$ and $y \in \mathbb{D}_m$ with $d\left(y, x\right) < \delta_n$ and $e\left(g_m\left(\omega\right)\left(y\right), g\left(x\right)\right) > \varepsilon$. Let $B_n = \cup_{\omega \in A} B_n(\omega)$. Then we can finish the proof along the lines of the proof of (ii).

Suppose Condition (b) holds. Repeat the proofs of (i), (ii), and (iii) under Condition (a) with the properties of g_n and g under Condition (b). For (ii), let $B_n(\omega)$ be the set of all x such that there is a $y \in \mathbb{D}_n(\omega)$ with $d\left(y, x\right) < \delta_n$ and $e\left(g_n\left(\omega\right)\left(y\right), g\left(x\right)\right) > \varepsilon$. For (iii), let $B_n(\omega)$ be the set of all x such that there are $m \geq n$ and $y \in \mathbb{D}_m(\omega)$ with $d\left(y, x\right) < \delta_n$ and $e\left(g_m\left(\omega\right)\left(y\right), g\left(x\right)\right) > \varepsilon$. The key difference is that Condition (a) requires that $X_n(\omega) \in \mathbb{D}_n$ for all ω holds for some fixed $\mathbb{D}_n$. Condition (b) only requires that $X_n(\omega) \in \mathbb{D}_n(\omega)$ for all ω holds for some random $\mathbb{D}_n$ which can take different values $\mathbb{D}_n(\omega)$ for different ω. On the other hand, Condition (b) strengthens the properties of g_n and g so that the claims hold as well. ■

Theorem A.2 (Extended delta method) *Let $\mathbb{D}$ and $\mathbb{E}$ be metrizable topological vector spaces, and let r_n be constants with $r_n \rightarrow \infty$. Let $\hat{\phi}_n : \Omega \rightarrow \mathbb{D}_{\mathcal{F}} \subset \mathbb{D}$ be a random element for every n.*

Let $\mathbb{D}_0 \subset \mathbb{D}$.

(i) Let $\mathcal{F} : \mathbb{D}_{\mathcal{F}} \to \mathbb{E}$ satisfy the condition that for every $\varepsilon > 0$, there is a measurable set $A \subset \Omega$ with $\mathbb{P}(A) \geq 1 - \varepsilon$ such that for some map $\mathcal{F}'_\phi$ on $\mathbb{D}_0$,

$$r_n(\mathcal{F}(\hat{\phi}_n + r_n^{-1} h_n) - \mathcal{F}(\hat{\phi}_n)) \to \mathcal{F}'_\phi(h)$$

*uniformly on A for every convergent sequence $\{h_n\} \subset \mathbb{D}$ with $\hat{\phi}_n(\omega) + r_n^{-1} h_n \in \mathbb{D}_{\mathcal{F}}$ for all n and all ω and $h_n \to h \in \mathbb{D}_0$. If $X_n : \Omega \to \mathbb{D}_{\mathcal{F}}$ are maps **with** $X_n(\omega) - \hat{\phi}_n(\omega) + \hat{\phi}_n(\omega') \in \mathbb{D}_{\mathcal{F}}$ for all $\omega, \omega' \in \Omega$ and $r_n(X_n - \hat{\phi}_n) \rightsquigarrow X$, where X is separable and takes its values in $\mathbb{D}_0$, then $r_n(\mathcal{F}(X_n) - \mathcal{F}(\hat{\phi}_n)) \rightsquigarrow \mathcal{F}'_\phi(X)$. Moreover, if $\mathcal{F}'_\phi$ is continuous on all of $\mathbb{D}$, then $r_n(\mathcal{F}(X_n) - \mathcal{F}(\hat{\phi}_n)) - \mathcal{F}'_\phi(r_n(X_n - \hat{\phi}_n))$ converges to zero in outer probability.*

*(ii) Let $\mathcal{F} : \mathbb{D}_{\mathcal{F}} \to \mathbb{E}$ satisfy the condition that for every $\varepsilon > 0$, there is a measurable set $A \subset \Omega$ with $\mathbb{P}(A) \geq 1 - \varepsilon$ such that for some **continuous** map $\mathcal{F}'_\phi$ on $\mathbb{D}_0$,*

$$r_{n_m}\{\mathcal{F}(\hat{\phi}_{n_m}(\omega_{n_m}) + r_{n_m}^{-1} h_{n_m}) - \mathcal{F}(\hat{\phi}_{n_m}(\omega_{n_m}))\} \to \mathcal{F}'_\phi(h)$$

for every convergent subsequence $\{h_{n_m}\} \subset \mathbb{D}$ with $\hat{\phi}_{n_m}(\omega_{n_m}) + r_{n_m}^{-1} h_{n_m} \in \mathbb{D}_{\mathcal{F}}$, $\omega_{n_m} \in A$, and $h_{n_m} \to h \in \mathbb{D}_0$. If $X_n : \Omega \to \mathbb{D}_{\mathcal{F}}$ are maps with $r_n(X_n - \hat{\phi}_n) \rightsquigarrow X$, where X takes its values in $\mathbb{D}_0$, then $r_n(\mathcal{F}(X_n) - \mathcal{F}(\hat{\phi}_n)) \rightsquigarrow \mathcal{F}'_\phi(X)$. Moreover, if $\mathcal{F}'_\phi$ is continuous on all of $\mathbb{D}$, then $r_n(\mathcal{F}(X_n) - \mathcal{F}(\hat{\phi}_n)) - \mathcal{F}'_\phi(r_n(X_n - \hat{\phi}_n))$ converges to zero in outer probability.

Remark A.2 *Theorem A.2 is an extension of Theorem 3.9.5 (delta method) of van der Vaart and Wellner (1996). Here, $\hat{\phi}_n$ is allowed to be random, which is the key difference between the two theorems. Theorem A.2 is used to establish the asymptotic distribution of the test statistic under null.*

Proof of Theorem A.2. (i). The proof mainly relies on the results of Theorem A.1. Define $\mathbb{D}_n(\omega) = \{h \in \mathbb{D} : \hat{\phi}_n(\omega) + r_n^{-1} h \in \mathbb{D}_{\mathcal{F}}\}$ for every n and every $\omega \in \Omega$. Let $\mathbb{D}_n = \cap_{\omega \in \Omega} \mathbb{D}_n(\omega)$. Define $g_n(\omega)(h) = r_n(\mathcal{F}(\hat{\phi}_n(\omega) + r_n^{-1} h) - \mathcal{F}(\hat{\phi}_n(\omega)))$ for every n, every $\omega \in \Omega$, and every $h \in \mathbb{D}_n$. Here, g_n is a random map because of $\hat{\phi}_n$. For every n and every $\omega \in \Omega$, $g_n(\omega) : \mathbb{D}_n \to \mathbb{E}$. By the assumptions, for every $\varepsilon > 0$ there is a measurable set $A \subset \Omega$ with $\mathbb{P}(A) \geq 1 - \varepsilon$ such that if $h_n \in \mathbb{D}_n$ with $h_n \to h \in \mathbb{D}_0$, then $g_n(h_n) \to \mathcal{F}'_\phi(h)$ uniformly on A. Also, $r_n(X_n(\omega) - \hat{\phi}_n(\omega)) \in \mathbb{D}_n$ for all ω by assumption. Now by Theorem A.1(i) (under Condition (a)),

$$r_n(\mathcal{F}(X_n) - \mathcal{F}(\hat{\phi}_n)) = g_n(r_n(X_n - \hat{\phi}_n)) \rightsquigarrow \mathcal{F}'_\phi(X).$$

Moreover, suppose $\mathcal{F}'_\phi$ is continuous on all of $\mathbb{D}$, and let $f_n(h) = (g_n(h), \mathcal{F}'_\phi(h))$ for every

$h \in \mathbb{D}_n$. By Theorem A.1(i) again,

$$\left(r_n(\mathcal{F}(X_n) - \mathcal{F}(\hat{\phi}_n)), \mathcal{F}'_\phi(r_n(X_n - \hat{\phi}_n))\right) = f_n(r_n(X_n - \hat{\phi}_n)) \rightsquigarrow \left(\mathcal{F}'_\phi, \mathcal{F}'_\phi\right)(X).$$

Thus by Theorem 1.3.6 (continuous mapping) of van der Vaart and Wellner (1996), $r_n(\mathcal{F}(X_n) - \mathcal{F}(\hat{\phi}_n)) - \mathcal{F}'_\phi(r_n(X_n - \hat{\phi}_n)) \rightsquigarrow 0$. The claim follows from Lemma 1.10.2(iii) of van der Vaart and Wellner (1996).

(ii). Together with the continuity of $\mathcal{F}'_\phi$, by arguments similar to the proof of (i), we can show that the claim holds by Theorem A.1(i) (under Condition (b)). ■

B Conditioning Covariates

In this section, we consider the case where conditioning covariates may be present, that is, the random assignment assumption holds conditional on some covariates. Suppose X is a conditioning covariate vector with dimension d_X, let $\mathcal{X}$ be the set of possible values of X, and let $\mathcal{X} = \{x_1, \ldots, x_L\}$.

First, consider the case introduced in Section 2.2 where the treatment and the instrument are both multivalued (and ordered). A testable implication with conditioning covariates is as follows.

Lemma B.1 *A testable implication of the conditional version of Assumption 2.2 is that*

$$\begin{aligned}
&\mathbb{P}(Y \in B, D = d_{\max}|Z = z_k, X = x_l) \leq \mathbb{P}(Y \in B, D = d_{\max}|Z = z_{k+1}, X = x_l)\\
&\textit{and } \mathbb{P}(Y \in B, D = d_{\min}|Z = z_k, X = x_l) \geq \mathbb{P}(Y \in B, D = d_{\min}|Z = z_{k+1}, X = x_l);\\
&\mathbb{P}(D \in C|Z = z_k, X = x_l) \geq \mathbb{P}(D \in C|Z = z_{k+1}, X = x_l) \qquad \text{(B.1)}
\end{aligned}$$

for all k with $1 \leq k \leq K-1$, all l with $1 \leq l \leq L$, all $B \in \mathcal{B}_\mathbb{R}$, and all $C = (-\infty, c]$ with $c \in \mathbb{R}$.

Suppose $d_{\min} = 0$ and $d_{\max} = 1$ without loss of generality. Define function spaces

$$\begin{aligned}
&\mathcal{G} = \left\{\left(1_{\mathbb{R}\times\mathbb{R}\times\{z_k\}\times\{x_l\}}, 1_{\mathbb{R}\times\mathbb{R}\times\{z_{k+1}\}\times\{x_l\}}\right) : k = 1, \ldots, K-1, l = 1, \ldots, L\right\},\\
&\mathcal{H}_1 = \left\{(-1)^d \cdot 1_{B\times\{d\}\times\mathbb{R}\times\mathbb{R}^{d_X}} : B \text{ is a closed interval}, d \in \{0,1\}\right\},\\
&\mathcal{H}_2 = \left\{1_{\mathbb{R}\times C\times\mathbb{R}\times\mathbb{R}^{d_X}} : C = (-\infty, c], c \in \mathbb{R}\right\}, \text{ and } \mathcal{H} = \mathcal{H}_1 \cup \mathcal{H}_2. \qquad \text{(B.2)}
\end{aligned}$$

For every probability measure Q with (12), we define ϕ_Q by $\phi_Q(h, g) = Q(h \cdot g_2)/Q(g_2) - Q(h \cdot g_1)/Q(g_1)$ for every $(h, g) \in \mathcal{H} \times \mathcal{G}$ with $g = (g_1, g_2)$. Testable implication (B.1) is equivalent to the H_0 in

$$H_0 : \sup_{(h,g)\in\mathcal{H}\times\mathcal{G}} \phi_Q(h,g) \leq 0 \text{ and } H_1 : \sup_{(h,g)\in\mathcal{H}\times\mathcal{G}} \phi_Q(h,g) > 0$$

if Q is the underlying probability distribution of the data. Then we can follow the test procedure in Section 3.1.1 to conduct the test with the function space $\mathcal{H}\times\mathcal{G}$ defined by the $\mathcal{H}$ and the $\mathcal{G}$ in (B.2).

Second, consider the case introduced in Section 2.3 where the treatment and the instrument can both be unordered. A testable implication with conditioning covariates is as follows.

Lemma B.2 *A testable implication of the conditional version of Assumption 2.4 is given by*

$$\mathbb{P}\left(Y \in B, D = d | Z = z', X = x_l\right) \leq \mathbb{P}\left(Y \in B, D = d | Z = z, X = x_l\right) \tag{B.3}$$

for all Borel sets B, all $(d, z, z') \in \mathcal{C}$, and all l with $1 \leq l \leq L$, where $\mathcal{C}$ is a prespecified subset of $\mathcal{D}\times\mathcal{Z}\times\mathcal{Z}$.

The inequality in (B.3) is similar to the generalized regression monotonicity (GRM) hypothesis in Hsu et al. (2019). The major difference is that Z is allowed to be unordered in (B.3). Define the function space

$$\mathcal{H}\times\mathcal{G} = \left\{ \begin{array}{c} \left(1_{B\times\{d\}\times\mathbb{R}\times\mathbb{R}^{d_X}}, \left(1_{\mathbb{R}\times\mathbb{R}\times\{z\}\times\{x_l\}}, 1_{\mathbb{R}\times\mathbb{R}\times\{z'\}\times\{x_l\}}\right)\right) : B \text{ is a closed interval,} \\ (d, z, z') \in \mathcal{C}, l = 1, \ldots, L \end{array} \right\}. \tag{B.4}$$

For every probability measure Q with (12), we define ϕ_Q by $\phi_Q(h,g) = Q(h\cdot g_2)/Q(g_2) - Q(h\cdot g_1)/Q(g_1)$ for every $(h,g) \in \mathcal{H}\times\mathcal{G}$ with $g = (g_1, g_2)$. Testable implication (B.3) is equivalent to the H_0 in

$$H_0 : \sup_{(h,g)\in\mathcal{H}\times\mathcal{G}} \phi_Q(h,g) \leq 0 \text{ and } H_1 : \sup_{(h,g)\in\mathcal{H}\times\mathcal{G}} \phi_Q(h,g) > 0$$

if Q is the underlying probability distribution of the data. Then we can follow the test procedure in Section 3.1.1 to conduct the test with the function space $\mathcal{H}\times\mathcal{G}$ defined in (B.4).

References

Abadie, A. (2002). Bootstrap tests for distributional treatment effects in instrumental variable models. *Journal of the American Statistical Association*, 97(457):284–292.

Abadie, A., Angrist, J., and Imbens, G. (2002). Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica*, 70(1):91–117.

Ananat, E. O. and Michaels, G. (2008). The effect of marital breakup on the income distribution of women with children. *Journal of Human Resources*, 43(3):611–629.

Andrews, D. W. (2000). Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica*, 68(2):399–405.

Andrews, D. W. and Shi, X. (2013). Inference based on conditional moment inequalities. *Econometrica*, 81(2):609–666.

Andrews, D. W. and Soares, G. (2010). Inference for parameters defined by moment inequalities using generalized moment selection. *Econometrica*, 78(1):119–157.

Angrist, J. D. (1990). Lifetime earnings and the Vietnam era draft lottery: Evidence from social security administrative records. *The American Economic Review*, 80(3):313–336.

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

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455.

Angrist, J. D. and Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, 106(4):979–1014.

Angrist, J. D. and Krueger, A. B. (1995). Split-sample instrumental variables estimates of the return to schooling. *Journal of Business & Economic Statistics*, 13(2):225–235.

Angrist, J. D. and Pischke, J.-S. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Angrist, J. D. and Pischke, J.-S. (2014). *Mastering Metrics: The Path from Cause to Effect*. Princeton University Press.

Armstrong, T. B. (2014). Weighted KS statistics for inference on conditional moment inequalities. *Journal of Econometrics*, 181(2):92–116.

Armstrong, T. B. and Chan, H. P. (2016). Multiscale adaptive inference on conditional moment inequalities. *Journal of Econometrics*, 194(1):24–43.

Balke, A. and Pearl, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439):1171–1176.

Barrett, G. F. and Donald, S. G. (2003). Consistent tests for stochastic dominance. *Econometrica*, 71(1):71–104.

Barrett, G. F., Donald, S. G., and Bhattacharya, D. (2014). Consistent nonparametric tests for Lorenz dominance. *Journal of Business & Economic Statistics*, 32(1):1–13.

Beare, B. K. and Fang, Z. (2017). Weak convergence of the least concave majorant of estimators for a concave distribution function. *Electronic Journal of Statistics*, 11(2):3841–3870.

Beare, B. K. and Moon, J.-M. (2015). Nonparametric tests of density ratio ordering. *Econometric Theory*, 31(3):471–492.

Beare, B. K. and Shi, X. (2019). An improved bootstrap test of density ratio ordering. *Econometrics and Statistics*, 10:9–26.

Bound, J., Jaeger, D. A., and Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association*, 90(430):443–450.

Buckles, K. S. and Hungerman, D. M. (2013). Season of birth and later outcomes: Old questions, new answers. *Review of Economics and Statistics*, 95(3):711–724.

Card, D. (1993). Using geographic variation in college proximity to estimate the return to schooling. Working Paper 4483, National Bureau of Economic Research.

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

Cawley, J. and Meyerhoefer, C. (2012). The medical care costs of obesity: An instrumental variables approach. *Journal of Health Economics*, 31(1):219–230.

Chernozhukov, V., Kim, W., Lee, S., and Rosen, A. M. (2015). Implementing intersection bounds in stata. *The Stata Journal*, 15(1):21–44.

Chernozhukov, V., Lee, S., and Rosen, A. M. (2013). Intersection bounds: Estimation and inference. *Econometrica*, 81(2):667–737.

Chetverikov, D. (2018). Adaptive tests of conditional moment inequalities. *Econometric Theory*, 34(1):186–227.

Davydov, Y. A., Lifshits, M. A., and Smorodina, N. V. (1998). *Local Properties of Distributions of Stochastic Functionals*, volume 173. American Mathematical Society.

Donald, S. G. and Hsu, Y.-C. (2016). Improving the power of tests of stochastic dominance. *Econometric Reviews*, 35(4):553–585.

Dümbgen, L. (1993). On nondifferentiable functions and the bootstrap. *Probability Theory and Related Fields*, 95(1):125–140.

Eren, O. and Ozbeklik, S. (2014). Who benefits from Job Corps? A distributional analysis of an active labor market program. *Journal of Applied Econometrics*, 29(4):586–611.

Fang, Z. and Santos, A. (2018). Inference on directionally differentiable functions. *The Review of Economic Studies*, 86(1):377–412.

Folland, G. B. (1999). *Real Analysis: Modern Techniques and Their Applications*. John Wiley & Sons.

Frölich, M. and Melly, B. (2013). Unconditional quantile treatment effects under endogeneity. *Journal of Business & Economic Statistics*, 31(3):346–357.

Ghosal, S. and Van der Vaart, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*, volume 44. Cambridge University Press.

Giacomini, R., Politis, D. N., and White, H. (2013). A warp-speed method for conducting Monte Carlo experiments involving bootstrap estimators. *Econometric Theory*, 29(3):567–589.

Hansen, B. E. (2017). Regression kink with an unknown threshold. *Journal of Business & Economic Statistics*, 35(2):228–240.

Heckman, J. J. and Pinto, R. (2018). Unordered monotonicity. *Econometrica*, 86(1):1–35.

Heckman, J. J., Urzua, S., and Vytlacil, E. (2006). Understanding instrumental variables in models with essential heterogeneity. *The Review of Economics and Statistics*, 88(3):389–432.

Heckman, J. J., Urzua, S., and Vytlacil, E. (2008). Instrumental variables in models with multiple outcomes: The general unordered case. *Annales d'Economie et de Statistique*, pages 151–174.

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

Heckman, J. J. and Vytlacil, E. J. (2007). Econometric evaluation of social programs, part ii: Using the marginal treatment effect to organize alternative econometric estimators to evaluate social programs, and to forecast their effects in new environments. In *Handbook of Econometrics*, pages 4875–5143. Amsterdam: Elsevier.

Hirano, K. and Porter, J. R. (2012). Impossibility results for nondifferentiable functionals. *Econometrica*, 80(4):1769–1790.

Hong, H. and Li, J. (2018). The numerical delta method. *Journal of Econometrics*, 206(2):379–394.

Horváth, L., Kokoszka, P., and Zitikis, R. (2006). Testing for stochastic dominance using the weighted McFadden-type statistic. *Journal of Econometrics*, 133(1):191–205.

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

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

Huber, M. and Wüthrich, K. (2018). Local average and quantile treatment effects under endogeneity: A review. *Journal of Econometric Methods*, 8(1).

Imbens, G. (2014). Instrumental variables: An econometrician's perspective. Technical report, National Bureau of Economic Research.

Imbens, G. W. and Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475.

Imbens, G. W. and Manski, C. F. (2004). Confidence intervals for partially identified parameters. *Econometrica*, 72(6):1845–1857.

Imbens, G. W. and Rubin, D. B. (1997). Estimating outcome distributions for compliers in instrumental variables models. *The Review of Economic Studies*, 64(4):555–574.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Kédagni, D. and Mourifié, I. (2020). Generalized instrumental inequalities: Testing the instrumental variable independence assumption. *Biometrika*, 107(3):661–675.

Kitagawa, T. (2015). A test for instrument validity. *Econometrica*, 83(5):2043–2063.

Koenker, R., Chernozhukov, V., He, X., and Peng, L. (2017). *Handbook of Quantile Regression*. CRC Press.

Lee, S. and Salanié, B. (2018). Identifying effects of multivalued treatments. *Econometrica*, 86(6):1939–1963.

Lee, S., Song, K., and Whang, Y.-J. (2013). Testing functional inequalities. *Journal of Econometrics*, 172(1):14–32.

Linton, O., Song, K., and Whang, Y.-J. (2010). An improved bootstrap test of stochastic dominance. *Journal of Econometrics*, 154(2):186–202.

Liu, S., Mourifié, I., and Wan, Y. (2020). Two-way exclusion restrictions in models with heterogeneous treatment effects. *The Econometrics Journal*, 23(3):345–362.

Marshall, J. (2016). Coarsening bias: How coarse treatment measurement upwardly biases instrumental variable estimates. *Political Analysis*, 24(2):157–171.

Melly, B. and Wüthrich, K. (2017). Local quantile treatment effects. In *Handbook of Quantile Regression*, pages 145–164. Chapman and Hall/CRC.

Mogstad, M., Torgovitsky, A., and Walters, C. R. (2020). Policy evaluation with multiple instrumental variables. NBER 27546.

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

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

Pollard, D. (1990). Empirical processes: Theory and applications. In *NSF-CBMS Regional Conference Series in Probability and Statistics*, pages i–86. JSTOR.

Reed, W. J. (2001). The Pareto, Zipf and other power laws. *Economics Letters*, 74(1):15–19.

Reed, W. J. (2003). The Pareto law of incomes—An explanation and an extension. *Physica A: Statistical Mechanics and Its Applications*, 319:469–486.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688.

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

Splawa-Neyman, J., Dabrowska, D. M., and Speed, T. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, pages 465–472.

Sun, Z. and Beare, B. K. (2021). Improved nonparametric bootstrap tests of Lorenz dominance. *Journal of Business & Economic Statistics*, 39(1):189–199.

Toda, A. A. (2012). The double power law in income distribution: Explanations and evidence. *Journal of Economic Behavior & Organization*, 84(1):364–381.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer.

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

Vytlacil, E. (2006). Ordered discrete-choice selection models and local average treatment effect assumptions: Equivalence, nonequivalence, and representation results. *The Review of Economics and Statistics*, 88(3):578–581.

Instrument Validity for Heterogeneous Causal Effects Online Supplementary Appendix

Zhenting Sun
China Center for Economic Research
National School of Development
Peking University
zhentingsun@nsd.pku.edu.cn

October 3, 2022

For the multivalued ordered treatment case, we assume $\mathcal{D} = \{d_1, d_2, \ldots\}$ in the proofs to obtain more general results. Assumption 2.2 with $\mathcal{D} = \{d_1, d_2, \ldots\}$ is

(i) Instrument Exclusion: For all $d \in \mathcal{D}$, $Y_{dz_1} = Y_{dz_2} = \cdots = Y_{dz_K}$ almost surely.

(ii) Random Assignment: The variable Z is jointly independent of $(\tilde{Y}, \tilde{D})$, where

$$\tilde{Y} = \left(Y_{d_1z_1}, \ldots, Y_{d_1z_K}, Y_{d_2z_1}, \ldots, Y_{d_2z_K}, \ldots\right) \text{ and } \tilde{D} = \left(D_{z_1}, \ldots, D_{z_K}\right).$$

(iii) Instrument Monotonicity: The potential treatment response variables satisfy $D_{z_{k+1}} \geq D_{z_k}$ almost surely for all $k \in \{1, 2, \ldots, K-1\}$.

Without loss of generality, we may assume that both $d_{\min}$ and $d_{\max}$ exist with $d_{\min} = 0$ and $d_{\max} = 1$ for simplicity. If $d_{\min}$ and $d_{\max}$ exist, we can always normalize $d_{\min}$ and $d_{\max}$ to 0 and 1, respectively. Then the function spaces defined in (11) can be used for $\mathcal{D} = \{d_1, d_2, \ldots\}$. All the results hold for $\mathcal{D} = \{d_1, \ldots, d_J\}$.

C Proofs of Main Results

We first introduce the following notation. For every $A \subset \bar{\mathcal{H}} \times \mathcal{G}$, define a map $\mathcal{S}_A : \ell^\infty\left(\Xi \times \bar{\mathcal{H}} \times \mathcal{G}\right) \to \ell^\infty(\Xi)$ by

$$\mathcal{S}_A(\psi)(\xi) = \sup_{(h,g)\in A} \psi(\xi, h, g)$$

for all $\psi \in \ell^\infty\left(\Xi \times \bar{\mathcal{H}} \times \mathcal{G}\right)$. For simplicity of notation, we will write $\mathcal{S}$ for $\mathcal{S}_{\bar{\mathcal{H}}\times\mathcal{G}}$. Define $\mathcal{M} : \ell^\infty(\bar{\mathcal{H}} \times \mathcal{G}) \to \ell^\infty(\Xi \times \bar{\mathcal{H}} \times \mathcal{G})$ by

$$\mathcal{M}(\varphi)(\xi, h, g) = \max\{\xi, \varphi(h, g)\} \tag{C.1}$$

for all $\varphi \in \ell^\infty(\bar{\mathcal{H}} \times \mathcal{G})$ and all $(\xi, h, g) \in \Xi \times \bar{\mathcal{H}} \times \mathcal{G}$. Note that for every finite sample set,

$$\mathcal{S}_{\mathcal{H}\times\mathcal{G}}(\hat{\phi}_{P_n}/\mathcal{M}(\hat{\sigma}_{P_n})) = \mathcal{S}(\hat{\phi}_{P_n}/\mathcal{M}(\hat{\sigma}_{P_n})). \tag{C.2}$$

Define a function $\mathcal{I} : L^1(\nu) \to \mathbb{R}$ by $\mathcal{I}(f) = \int_\Xi f \, \mathrm{d}\nu$ for all $f \in L^1(\nu)$. Now we can write the test statistic in (24) as

$$\sqrt{T_n}\mathcal{I} \circ \mathcal{S}_{\mathcal{H}\times\mathcal{G}}\left(\frac{\hat{\phi}_{P_n}}{\mathcal{M}(\hat{\sigma}_{P_n})}\right). \tag{C.3}$$

Lemma C.1 *Let $\mathcal{P}$ be the set of probability measures defined in Section 3. Let $\mathcal{H}_1$, $\bar{\mathcal{H}}_1$, $\mathcal{H}_2$, $\bar{\mathcal{H}}_2$, $\mathcal{H}$, and $\bar{\mathcal{H}}$ be as in (11). Then for every $Q \in \mathcal{P}$, the closures of $\mathcal{H}_1$ and $\mathcal{H}_2$ in $L^2(Q)$ are equal to $\bar{\mathcal{H}}_1$ and $\bar{\mathcal{H}}_2$, respectively. Also, the closure of $\mathcal{H}$ in $L^2(Q)$ is equal to $\bar{\mathcal{H}}$ for every $Q \in \mathcal{P}$.*

Proof of Lemma C.1. Let $\mathcal{H}_{1d} = \{(-1)^d \cdot 1_{B\times\{d\}\times\mathbb{R}} : B \text{ is a closed interval in } \mathbb{R}\}$ for $d \in \{0, 1\}$. We first show that the closure of $\mathcal{H}_{1d}$ in $L^2(Q)$ is equal to

$$\bar{\mathcal{H}}_{1d} = \left\{(-1)^d \cdot 1_{B\times\{d\}\times\mathbb{R}} : B \text{ is a closed, open, or half-closed interval in } \mathbb{R}\right\}.$$

If this is true, the first claim of the Lemma follows from $\bar{\mathcal{H}}_1 = \bar{\mathcal{H}}_{10} \cup \bar{\mathcal{H}}_{11}$.

Suppose there is a sequence $\{h_n\} \subset \mathcal{H}_{1d}$ such that $\|h_n - h\|_{L^2(Q)} \to 0$ for some $h \in L^2(Q)$. Then h_n is a Cauchy sequence, that is, $\|h_n - h_m\|_{L^2(Q)} \to 0$ as $n, m \to \infty$. By the definition of $\mathcal{H}_{1d}$, $h_n = (-1)^d \cdot 1_{B_n\times\{d\}\times\mathbb{R}}$, where B_n is a closed interval in $\mathbb{R}$. It is possible that $\int 1_{B_n\times\{d\}\times\mathbb{R}} \, \mathrm{d}Q \to 0$, and in this case there is a $B = \{a\}$ for some $a \in \mathbb{R}$ such that $Q(B \times \mathbb{R} \times \mathbb{R}) = 0$ and $h_n \to (-1)^d \cdot 1_{B\times\{d\}\times\mathbb{R}} \in \mathcal{H}_{1d}$. If $\int 1_{B_n\times\{d\}\times\mathbb{R}} \, \mathrm{d}Q \not\to 0$, then there is an $\varepsilon > 0$ such that for all $n_\varepsilon > 0$, there is an $n > n_\varepsilon$ such that $\|h_n\|^2_{L^2(Q)} > \varepsilon$. For a $\delta_1 \ll \varepsilon$, there is an N_1 such that $\|h_n - h_m\|^2_{L^2(Q)} < \delta_1$ for all $m, n > N_1$. Thus there is an $n_1 > N_1$ such that $\|h_{n_1}\|^2_{L^2(Q)} > \varepsilon$ and $\|h_n - h_{n_1}\|^2_{L^2(Q)} < \delta_1$ for all $n > N_1$. Now let δ_2 be such that $0 < \delta_2 \ll \delta_1$. Then there is an $N_2 > n_1$ such that $\|h_n - h_m\|^2_{L^2(Q)} < \delta_2$ for all $m, n > N_2$. Thus there is an $n_2 > N_2$ such that $\|h_{n_2}\|^2_{L^2(Q)} > \varepsilon$ and $\|h_n - h_{n_2}\|^2_{L^2(Q)} < \delta_2$ for all $n > N_2$. In this way, we can find a sequence $\{h_{n_k}\}_k$ with $h_{n_k} = (-1)^d \cdot 1_{B_{n_k}\times\{d\}\times\mathbb{R}}$, $\|h_{n_k}\|^2_{L^2(Q)} > \varepsilon$, $\|h_n - h_{n_k}\|^2_{L^2(Q)} < \delta_k$ for all $n > n_k$, and $\delta_k \downarrow 0$. Let $B^\infty = \cup_{j=1}^\infty \cap_{k=j}^\infty B_{n_k}$. For every K, $\|h_{n_k} - h_{n_K}\|^2_{L^2(Q)} < \delta_K$ for all $k > K$. Notice that for every $K' > K$,

$$\|h_{n_K} - (-1)^d \cdot 1_{(\cap_{k=K'}^\infty B_{n_k})\times\{d\}\times\mathbb{R}}\|^2_{L^2(Q)} = \int |1_{B_{n_K}\times\{d\}\times\mathbb{R}} - 1_{(\cap_{k=K'}^\infty B_{n_k})\times\{d\}\times\mathbb{R}}|^2 \, \mathrm{d}Q$$
$$= \int 1_{B_{n_K}\setminus(\cap_{k=K'}^\infty B_{n_k})\times\{d\}\times\mathbb{R}} \, \mathrm{d}Q + \int 1_{(\cap_{k=K'}^\infty B_{n_k})\setminus B_{n_K}\times\{d\}\times\mathbb{R}} \, \mathrm{d}Q.$$

Because B_{n_k} is a closed interval for all k, we have that for every $K'' \geq K'$, there exist L_1 and L_2 with $K' \leq L_1 \leq L_2 \leq K''$ such that $\cup_{k=K'}^{K''}(B_{n_K} \setminus B_{n_k}) = (B_{n_K} \setminus B_{n_{L_1}}) \cup (B_{n_K} \setminus B_{n_{L_2}})$.

Then since

$$\|h_{n_k} - h_{n_K}\|_{L^2(Q)}^2 = Q(B_{n_K} \setminus B_{n_k} \times \{d\} \times \mathbb{R}) + Q(B_{n_k} \setminus B_{n_K} \times \{d\} \times \mathbb{R}) < \delta_K$$

for all $k > K$, we have

$$\begin{aligned}\int 1_{B_{n_K} \setminus (\cap_{k=K'}^{\infty} B_{n_k}) \times \{d\} \times \mathbb{R}} \, \mathrm{d}Q &= Q(B_{n_K} \setminus (\cap_{k=K'}^{\infty} B_{n_k}) \times \{d\} \times \mathbb{R}) \\ &= Q(\cup_{k=K'}^{\infty} (B_{n_K} \setminus B_{n_k}) \times \{d\} \times \mathbb{R}) \leq 2\delta_K.\end{aligned}$$

Similarly, it is easy to show that $\int 1_{(\cap_{k=K'}^{\infty} B_{n_k}) \setminus B_{n_K} \times \{d\} \times \mathbb{R}} \, \mathrm{d}Q \leq 2\delta_K$. Thus it follows that

$$\|h_{n_K} - (-1)^d \cdot 1_{(\cap_{k=K'}^{\infty} B_{n_k}) \times \{d\} \times \mathbb{R}}\|_{L^2(Q)}^2 \leq 4\delta_K,$$

which is true for all $K' > K$. Letting $K' \to \infty$, by the dominated convergence theorem ($B^{\infty} = \cup_{j=1}^{\infty} \cap_{k=j}^{\infty} B_{n_k}$) we have

$$\|h_{n_K} - (-1)^d \cdot 1_{B^{\infty} \times \{d\} \times \mathbb{R}}\|_{L^2(Q)}^2 \leq 4\delta_K.$$

This implies that $\|h_{n_K} - (-1)^d \cdot 1_{B^{\infty} \times \{d\} \times \mathbb{R}}\|_{L^2(Q)} \to 0$ as $K \to \infty$, because $\delta_K \downarrow 0$. Finally, we have

$$\|h_n - (-1)^d \cdot 1_{B^{\infty} \times \{d\} \times \mathbb{R}}\|_{L^2(Q)} \leq \|h_n - h_{n_K}\|_{L^2(Q)} + \|h_{n_K} - (-1)^d \cdot 1_{B^{\infty} \times \{d\} \times \mathbb{R}}\|_{L^2(Q)} \to 0.$$

Clearly, B^{∞} can be a closed, open, or half-closed interval in $\mathbb{R}$. Also, every element of $\bar{\mathcal{H}}_{1d}$ is equal to the limit of a sequence of elements of $\mathcal{H}_{1d}$ under the $L^2(Q)$ norm. Thus the closure of $\mathcal{H}_{1d}$ in $L^2(Q)$ is equal to $\bar{\mathcal{H}}_{1d}$ for every $Q \in \mathcal{P}$. Similarly, we can show that the closure of $\mathcal{H}_2$ in $L^2(Q)$ is equal to $\bar{\mathcal{H}}_2$ for every $Q \in \mathcal{P}$. As a result, the closure of $\mathcal{H} = \mathcal{H}_1 \cup \mathcal{H}_2$ in $L^2(Q)$ is equal to $\bar{\mathcal{H}} = \bar{\mathcal{H}}_1 \cup \bar{\mathcal{H}}_2$ for every $Q \in \mathcal{P}$. ■

Lemma C.2 *Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be defined as in (11). Then $\mathcal{H}_1$ is a VC class*[1] *with VC index $V(\mathcal{H}_1) = 3$, and $\mathcal{H}_2$ is a VC class with VC index $V(\mathcal{H}_2) = 2$.*

Proof of Lemma C.2. All the functions $h \in \mathcal{H}_1$ take the form $h = -1_{B \times \{1\} \times \mathbb{R}}$ or $h = 1_{B \times \{0\} \times \mathbb{R}}$, where B is a closed interval. If $h = -1_{B \times \{1\} \times \mathbb{R}}$, the subgraph of h is

$$C_{1B} = \left\{(y, w, z, t) \subset \mathbb{R}^4 : t < -1_{B \times \{1\} \times \mathbb{R}}(y, w, z)\right\}.$$

[1] See the definition of VC class of functions in van der Vaart and Wellner (1996, p. 141).

If $h = 1_{B\times\{0\}\times\mathbb{R}}$, the subgraph of h is

$$C_{0B} = \left\{(y, w, z, t) \subset \mathbb{R}^4 : t < 1_{B\times\{0\}\times\mathbb{R}}(y, w, z)\right\}.$$

Let $\mathcal{C} = \{C_{dB} : B \text{ is a closed interval in } \mathbb{R}, d \in \{0, 1\}\}$.

Suppose there are two different points $a_1 = (y_1, w_1, z_1, t_1), a_2 = (y_2, w_2, z_2, t_2) \in \mathbb{R}^4$ with $y_1 < y_2$, $w_1 = w_2 = 0$, and $0 \le t_1, t_2 < 1$. Then there is a point $\bar{y} \in (y_1, y_2)$. Let $B_0 = \{\bar{y}\}$, $B_1 = [y_1, \bar{y}]$, $B_2 = [\bar{y}, y_2]$, and $B_3 = [y_1, y_2]$. Now we have $\varnothing = C_{0B_0} \cap \{a_1, a_2\}$, $\{a_1\} = C_{0B_1} \cap \{a_1, a_2\}$, $\{a_2\} = C_{0B_2} \cap \{a_1, a_2\}$, and $\{a_1, a_2\} = C_{0B_3} \cap \{a_1, a_2\}$. Thus $\mathcal{C}$ shatters $\{a_1, a_2\}$.

Suppose now there are three different points $a_1 = (y_1, w_1, z_1, t_1)$, $a_2 = (y_2, w_2, z_2, t_2)$, $a_3 = (y_3, w_3, z_3, t_3)$ in $\mathbb{R}^4$. Without loss of generality, suppose $t_1 \le t_2 \le t_3 < 1$, so that it is possible for $\mathcal{C}$ to pick out $\{a_j\}$ for each $j \in \{1, 2, 3\}$.

(1) Suppose $t_1 \ge 0$. In this case, we need $w_1 = w_2 = w_3 = 0$ in order to pick out $\{a_j\}$ for each j. Without loss of generality, suppose $y_1 \le y_2 \le y_3$. If we want $\mathcal{C}$ to pick out $\{a_1, a_3\}$, we need to find a closed interval B such that $y_1, y_3 \in B$, in which case $a_1, a_3 \in C_{0B}$. However, $a_2 \in C_{0B}$ for all such B.

(2) Suppose $t_1 < 0$, $t_2 \ge 0$. Then we need $w_2 = w_3 = 0$ in order to pick out $\{a_j\}$ for each $j \in \{2, 3\}$ by using C_{0B} for some closed interval B. But in this case, $\mathcal{C}$ can never pick out $\{a_2\}$, $\{a_3\}$, or $\{a_2, a_3\}$, since for every closed interval B, $a_1 \in C_{0B}$.

(3) Suppose $t_1, t_2 < 0$, $t_3 \ge 0$. Then we need $w_3 = 0$ in order to pick out $\{a_3\}$ by using C_{0B} for some closed interval B. In this case, $\mathcal{C}$ can never pick out $\{a_3\}$, since for every closed interval B, $a_1, a_2 \in C_{0B}$.

(4) Suppose $t_1, t_2, t_3 < 0$. Then for every closed interval B, $a_1, a_2, a_3 \in C_{0B}$. If we want $\mathcal{C}$ to pick out $\{a_j, a_{j'}\}$ for all $j \ne j'$, we need to use C_{1B}. If $w_j \ne 1$, then for every B, $a_j \in C_{1B}$. Thus we consider $w_1 = w_2 = w_3 = 1$.

  (a) Suppose $-1 \le t_1, t_2, t_3 < 0$. Without loss of generality, we assume that $y_1 \le y_2 \le y_3$. But now if we want $\mathcal{C}$ to pick out $\{a_2\}$, we need to find a closed interval B such that $y_1, y_3 \in B$ but $y_2 \notin B$, which is not possible.

  (b) Suppose $t_j < -1$ for some $j \in \{1, 2, 3\}$. In this case, $a_j \in C_{1B}$ for every closed interval B.

Therefore, we conclude that $\mathcal{H}_1$ is a VC class with VC index $V(\mathcal{H}_1) = 3$. Similarly, we can show that $\mathcal{H}_2$ is a VC class with VC index $V(\mathcal{H}_2) = 2$. ■

Lemma C.3 *Let $\mathcal{H}$ be defined as in (11). Then $\mathcal{H}$ is totally bounded under $\|\cdot\|_{L^r(Q)}$ for every probability measure $Q \in \mathcal{P}$ and every $r \geq 1$.*

Proof of Lemma C.3. Let $N(\varepsilon, \mathcal{H}_j, L^r(Q))$ denote the covering number under the $L^r(Q)$ norm for $\mathcal{H}_j$ for $j \in \{1, 2\}$ and all $\varepsilon > 0$, where $\mathcal{H}_j$ is defined as in (11). Since $\mathcal{H}_1$ and $\mathcal{H}_2$ are VC classes by Lemma C.2 with $V(\mathcal{H}_1) = 3$ and $V(\mathcal{H}_2) = 2$, by Theorem 2.6.7 of van der Vaart and Wellner (1996) with envelope function $F = 1$ and $r \geq 1$ we have that for every probability measure Q,

$$N(\varepsilon, \mathcal{H}_1, L^r(Q)) \leq K_1 3 (16e)^3 (1/\varepsilon)^{2r} \text{ and } N(\varepsilon, \mathcal{H}_2, L^r(Q)) \leq K_2 2 (16e)^2 (1/\varepsilon)^r$$

for universal constants $K_1, K_2 \geq 1$ and every $\varepsilon \in (0, 1)$. Since $\mathcal{H} = \mathcal{H}_1 \cup \mathcal{H}_2$, we have

$$N(\varepsilon, \mathcal{H}, L^r(Q)) \leq K_1 3 (16e)^3 (1/\varepsilon)^{2r} + K_2 2 (16e)^2 (1/\varepsilon)^r, \tag{C.4}$$

which implies that $\mathcal{H}$ is totally bounded. ■

Lemma C.4 *Let $\bar{\mathcal{H}}$ be as in (11). Then $\bar{\mathcal{H}}$ is compact under $\|\cdot\|_{L^2(Q)}$ for every $Q \in \mathcal{P}$.*

Proof of Lemma C.4. By Lemma C.3, $\mathcal{H}$ is totally bounded under $\|\cdot\|_{L^2(Q)}$ for all $Q \in \mathcal{P}$. Suppose that $\mathcal{H} \subset \bigcup_{j \in J} B_{\varepsilon/2}(h_j)$, where J is a finite index set and $B_{\varepsilon/2}(h_j)$ is an open ball with center h_j and radius $\varepsilon/2$ under $\|\cdot\|_{L^2(Q)}$. By Lemma C.1, $\bar{\mathcal{H}}$ is equal to the closure of $\mathcal{H}$ in $L^2(Q)$. Clearly, $\bar{\mathcal{H}} \subset \bigcup_{j \in J} \overline{B_{\varepsilon/2}(h_j)} \subset \bigcup_{j \in J} B_\varepsilon(h_j)$, and therefore

$$N(\varepsilon, \bar{\mathcal{H}}, L^2(Q)) \leq N(\varepsilon/2, \mathcal{H}, L^2(Q)), \tag{C.5}$$

which, together with (C.4), implies that $\bar{\mathcal{H}}$ is totally bounded. Since $L^2(Q)$ is complete, $\bar{\mathcal{H}}$ is compact in $L^2(Q)$. ■

Let $\bar{\mathcal{H}}$ and $\mathcal{G}_K$ be defined as in (11). Let $\mathcal{V} = \{h \cdot f : h \in \bar{\mathcal{H}}, f \in \mathcal{G}_K\}$. Then define

$$\tilde{\mathcal{V}} = \mathcal{V} \cup \mathcal{G}_K. \tag{C.6}$$

Lemma C.5 *The function space $\tilde{\mathcal{V}}$ is Donsker and pre-Gaussian uniformly in $Q \in \mathcal{P}$.*

Proof of Lemma C.5. For every $\delta > 0$ and every $Q \in \mathcal{P}$, define

$$\tilde{\mathcal{V}}_{\delta,Q} = \left\{v - v' : v, v' \in \tilde{\mathcal{V}}, \|v - v'\|_{L^2(Q)} < \delta\right\} \text{ and } \tilde{\mathcal{V}}^2_\infty = \left\{(v - v')^2 : v, v' \in \tilde{\mathcal{V}}\right\}.$$

First, we show that $\tilde{\mathcal{V}}_{\delta,Q}$ is Q-measurable[2] for all $Q \in \mathcal{P}$. Similar to the construction of $\mathcal{H}$,

[2] See Definition 2.3.3 of Q-measurable class in van der Vaart and Wellner (1996).

we construct function spaces by

$$\mathcal{H}_{q1} = \left\{ (-1)^d \cdot 1_{B\times\{d\}\times\mathbb{R}} : B = [a,b]\,, a,b \in \mathbb{Q}, a \le b, d \in \{0,1\} \right\},$$
$$\mathcal{H}_{q2} = \{ 1_{\mathbb{R}\times C\times\mathbb{R}} : C = (-\infty, c]\,, c \in \mathbb{Q} \}, \text{ and } \mathcal{H}_q = \mathcal{H}_{q1} \cup \mathcal{H}_{q2},$$

where $\mathbb{Q}$ denotes the set of all rational numbers. Now define

$$\tilde{\mathcal{V}}_q = \{ h \cdot f : h \in \mathcal{H}_q, f \in \mathcal{G}_K \} \cup \mathcal{G}_K \text{ and } \tilde{\mathcal{V}}_{q\delta,Q} = \left\{ v - v' : v, v' \in \tilde{\mathcal{V}}_q, \left\| v - v' \right\|_{L^2(Q)} < \delta \right\}.$$

By construction, $\mathcal{G}_K$ is a finite set. Since $\mathbb{Q}$ is countable (and therefore the set of ordered pairs of elements of $\mathbb{Q}$ is countable), $\mathcal{H}_{q1}$ and $\mathcal{H}_{q2}$ are countable (and therefore $\mathcal{H}_q$ and $\tilde{\mathcal{V}}_q$ are countable).

Clearly, $\tilde{\mathcal{V}}_{q\delta,Q}$ is a countable subset of $\tilde{\mathcal{V}}_{\delta,Q}$. For every $v \in \tilde{\mathcal{V}}$, there is a sequence $\{v_m\} \subset \tilde{\mathcal{V}}_q$ such that $v_m \to v$ pointwise, because $\mathbb{Q}$ is dense in $\mathbb{R}$. For example, if $v = (-1)^d \cdot 1_{(\sqrt{2},\sqrt{3}]\times\{d\}\times\mathbb{R}} \cdot 1_{\mathbb{R}\times\mathbb{R}\times\{z_k\}}$, we can find $v_m = (-1)^d \cdot 1_{[a_m,b_m]\times\{d\}\times\mathbb{R}} \cdot 1_{\mathbb{R}\times\mathbb{R}\times\{z_k\}}$ with $a_m \downarrow \sqrt{2}$, $b_m \downarrow \sqrt{3}$, and $a_m, b_m \in \mathbb{Q}$. Suppose $v - v' \in \tilde{\mathcal{V}}_{\delta,Q}$ and $v_m, v'_m \in \tilde{\mathcal{V}}_q$ such that $v_m \to v$ and $v'_m \to v'$ pointwise. It is easy to show that $\|v_m - v'_m\|_{L^2(Q)} < \delta$ for large m, that is, $v_m - v'_m \in \tilde{\mathcal{V}}_{q\delta,Q}$ for large m. By Example 2.3.4 of van der Vaart and Wellner (1996), $\tilde{\mathcal{V}}_{\delta,Q}$ is Q-measurable, and this is true for all $\delta > 0$. Similarly, $\tilde{\mathcal{V}}^2_\infty$ is Q-measurable.

By the construction of $\tilde{\mathcal{V}}$, $F = 1$ is a measurable envelope function with $\int F^2 \,\mathrm{d}Q < \infty$. Also, $\lim_{M\to\infty} \sup_{Q\in\mathcal{P}} \int F^2 \cdot 1\{F > M\}\,\mathrm{d}Q = 0$. For all $H \in \mathcal{P}$ and all $\varepsilon \ge 2$,

$$N\left(\varepsilon \|F\|_{L^2(H)}, \tilde{\mathcal{V}}, L^2(H)\right) = N\left(\varepsilon, \tilde{\mathcal{V}}, L^2(H)\right) = 1. \tag{C.7}$$

For all $H \in \mathcal{P}$ and all $\varepsilon > 0$,

$$N\left(\varepsilon, \mathcal{V}, L^2(H)\right) \le N\left(\frac{\varepsilon}{2}, \bar{\mathcal{H}}, L^2(H)\right) \cdot N\left(\frac{\varepsilon}{2}, \mathcal{G}_K, L^2(H)\right) \le K \cdot N\left(\frac{\varepsilon}{2}, \bar{\mathcal{H}}, L^2(H)\right), \tag{C.8}$$

where K is the number of elements in $\mathcal{G}_K$. Thus by the definition of $\tilde{\mathcal{V}}$ in (C.6),

$$N\left(\varepsilon, \tilde{\mathcal{V}}, L^2(H)\right) \le K \cdot N\left(\frac{\varepsilon}{2}, \bar{\mathcal{H}}, L^2(H)\right) + K \tag{C.9}$$

for all $H \in \mathcal{P}$ and all $\varepsilon > 0$. Let $\mathcal{Q}$ denote the set of finitely discrete probability measures. The results in (C.4), (C.5), (C.7), and (C.9) imply that

$$\int_0^\infty \sup_{H\in\mathcal{Q}} \sqrt{\log N\left(\varepsilon \|F\|_{L^2(H)}, \tilde{\mathcal{V}}, L^2(H)\right)}\,\mathrm{d}\varepsilon = \int_0^2 \sup_{H\in\mathcal{Q}} \sqrt{\log N\left(\varepsilon, \tilde{\mathcal{V}}, L^2(H)\right)}\,\mathrm{d}\varepsilon$$
$$\le \int_0^2 \sqrt{\log\left\{K \cdot (K_1 + K_2) \cdot 3 \cdot (16e)^3 (4/\varepsilon)^4 + K\right\}}\,\mathrm{d}\varepsilon < \infty.$$

The claim of the Lemma follows from Theorem 2.8.3 of van der Vaart and Wellner (1996). ■

Lemma C.6 *The function space $\tilde{\mathcal{V}}$ defined in (C.6) is Glivenko–Cantelli uniformly in $Q \in \mathcal{P}$.*

Proof of Lemma C.6. Similar to the proof of Lemma C.5, we can show that $\tilde{\mathcal{V}}$ is Q-measurable for every $Q \in \mathcal{P}$. With $F = 1$ being an envelope function of $\tilde{\mathcal{V}}$, we have $\lim_{M\to\infty} \sup_{Q\in\mathcal{P}} \int F \cdot 1\{F > M\}\, \mathrm{d}Q = 0$. Similar to the proofs of Lemmas C.1, C.4, and C.5, we can show that for every $Q \in \mathcal{P}$ and every $\varepsilon > 0$, the closure of $\mathcal{H}$ in $L^1(Q)$ is equal to $\bar{\mathcal{H}}$, $N(\varepsilon, \bar{\mathcal{H}}, L^1(Q)) \leq N(\varepsilon/2, \mathcal{H}, L^1(Q))$, and $N(\varepsilon, \tilde{\mathcal{V}}, L^1(Q)) \leq K \cdot N\left(\varepsilon/2, \bar{\mathcal{H}}, L^1(Q)\right) + K$. Then by (C.4), we can show that $\sup_{H\in\mathcal{Q}_n} \log N(\varepsilon \left\|F\right\|_{L^1(H)}, \tilde{\mathcal{V}}, L^1(H)) = o(n)$ with the envelope function $F = 1$, where $\mathcal{Q}_n$ is the collection of all possible realizations of empirical measures of n observations. Then by Theorem 2.8.1 in van der Vaart and Wellner (1996), $\tilde{\mathcal{V}}$ is Glivenko–Cantelli uniformly in $Q \in \mathcal{P}$. ■

Lemma C.7 *Let $\mathcal{H}$ and $\mathcal{G}$ be defined as in (11), let ρ_P be as in (18), and define $\overline{\mathcal{H}\times\mathcal{G}}$ as the closure of $\mathcal{H}\times\mathcal{G}$ in $L^2(P)\times(L^2(P)\times L^2(P))$ under ρ_P. Then $N\left(\varepsilon, \overline{\mathcal{H}\times\mathcal{G}}, \rho_P\right) = O\left(1/\varepsilon^4\right)$ as $\varepsilon \to 0$.*

Proof of Lemma C.7. By the constructions of $\mathcal{H}\times\mathcal{G}$ and the metric ρ_P,

$$N\left(\varepsilon, \mathcal{H}\times\mathcal{G}, \rho_P\right) \leq N\left(\frac{\varepsilon}{3}, \mathcal{H}, L^2(P)\right) \cdot \left[N\left(\frac{\varepsilon}{3}, \mathcal{G}_K, L^2(P)\right)\right]^2,$$

where $\mathcal{G}_K$ is defined as in (11). By the construction of $\mathcal{G}_K$, $N\left(\varepsilon/3, \mathcal{G}_K, L^2(P)\right) \leq K$, where K is the number of elements in $\mathcal{G}_K$. This, together with (C.4), implies that $N\left(\varepsilon, \mathcal{H}\times\mathcal{G}, \rho_P\right) = O\left(1/\varepsilon^4\right)$ as $\varepsilon \to 0$. Similar to (C.5),

$$N\left(\varepsilon, \overline{\mathcal{H}\times\mathcal{G}}, \rho_P\right) \leq N\left(\frac{\varepsilon}{2}, \mathcal{H}\times\mathcal{G}, \rho_P\right) = O\left(\frac{1}{\varepsilon^4}\right) \text{ as } \varepsilon \to 0.$$

■

Lemma C.8 *Let $\mathcal{H}$ and $\mathcal{G}$ be defined as in (11), and let ρ_P be as in (18). Then $\overline{\mathcal{H}\times\mathcal{G}}$, the closure of $\mathcal{H}\times\mathcal{G}$ under ρ_P in Lemma C.7, is compact and $\overline{\mathcal{H}\times\mathcal{G}} = \bar{\mathcal{H}}\times\mathcal{G}$, where $\bar{\mathcal{H}}$ is defined as in (11).*

Proof of Lemma C.8. The first claim follows from Lemma C.7 and the fact that $L^2(P)\times(L^2(P)\times L^2(P))$ is complete under ρ_P. The second claim holds by the constructions of ρ_P and $\mathcal{G}$. ■

Proof of Lemma 2.1. Suppose Assumption 2.2 holds with $\mathcal{D} = \{d_1, d_2, \ldots\}$. Then we can define Y_d by $Y_d = Y_{dz_1} = Y_{dz_2} = \cdots = Y_{dz_K}$ almost surely for all $d \in \mathcal{D}$. First, suppose $d_{\max}$ exists. Under Assumption 2.2, for all k with $1 \leq k \leq K-1$ and all Borel sets B,

$$\begin{aligned}&\mathbb{P}\left(Y \in B, D = d_{\max} | Z = z_k\right) = \mathbb{P}\left(Y_{d_{\max}} \in B, D_{z_k} = d_{\max}\right)\\&= \sum_j \mathbb{P}\left(Y_{d_{\max}} \in B, D_{z_k} = d_{\max}, D_{z_{k+1}} = d_j\right) = \mathbb{P}\left(Y_{d_{\max}} \in B, D_{z_k} = d_{\max}, D_{z_{k+1}} = d_{\max}\right)\end{aligned}$$

and

$$\begin{aligned}\mathbb{P}\left(Y \in B, D = d_{\max} | Z = z_{k+1}\right) &= \mathbb{P}\left(Y_{d_{\max}} \in B, D_{z_{k+1}} = d_{\max}\right)\\&= \sum_j \mathbb{P}\left(Y_{d_{\max}} \in B, D_{z_k} = d_j, D_{z_{k+1}} = d_{\max}\right).\end{aligned}$$

Thus $\mathbb{P}\left(Y \in B, D = d_{\max} | Z = z_{k+1}\right) \geq \mathbb{P}\left(Y \in B, D = d_{\max} | Z = z_k\right)$. Second, suppose $d_{\min}$ exists. Then similarly, $\mathbb{P}\left(Y \in B, D = d_{\min} | Z = z_k\right) \geq \mathbb{P}\left(Y \in B, D = d_{\min} | Z = z_{k+1}\right)$. ■

Remark C.1 *Lemma B.1 can be proved analogously. The proofs of Lemmas 2.2 and B.2 are trivial.*

Lemma C.9 *Let $\mathbb{D}_{\mathcal{L}} = \{R \in \ell^\infty(\tilde{\mathcal{V}}) : R(h \cdot g_l)/R(g_l)$ exists for all $h \in \bar{\mathcal{H}}$ and all $g_l \in \mathcal{G}_K\}$. Define $\mathcal{L} : \mathbb{D}_{\mathcal{L}} \subset \ell^\infty(\tilde{\mathcal{V}}) \to \ell^\infty\left(\bar{\mathcal{H}} \times \mathcal{G}\right)$ by*

$$\mathcal{L}(R)(h, g) = \frac{R(h \cdot g_2)}{R(g_2)} - \frac{R(h \cdot g_1)}{R(g_1)}$$

for all $R \in \mathbb{D}_{\mathcal{L}}$ and all $(h, g) \in \bar{\mathcal{H}} \times \mathcal{G}$ with $g = (g_1, g_2)$. Then $\mathcal{L}$ is uniformly Hadamard differentiable[3] along every sequence $P_n \to P$ in $\mathbb{D}_{\mathcal{L}}$, tangentially to $\ell^\infty(\tilde{\mathcal{V}})$, with the derivative $\mathcal{L}'_P$ defined by

$$\mathcal{L}'_P(H)(h, g) = \frac{H(h \cdot g_2) P(g_2) - P(h \cdot g_2) H(g_2)}{P^2(g_2)} - \frac{H(h \cdot g_1) P(g_1) - P(h \cdot g_1) H(g_1)}{P^2(g_1)} \tag{C.10}$$

for all $H \in \ell^\infty(\tilde{\mathcal{V}})$.[4]

Remark C.2 *By the definition of $\mathcal{L}$, $\mathcal{L}(Q) = \phi_Q$ for all $Q \in \mathcal{P}$. We will apply Lemma C.9 along with the suitable delta method to deduce the asymptotic distributions of $\sqrt{n}(\hat{\phi}_{P_n} - \phi_P)$ and the bootstrap version of this random element.*

[3]See the definitions of Hadamard differentiability and uniform Hadamard differentiability in van der Vaart and Wellner (1996, pp. 372–375).

[4]By (13), $\mathcal{L}'_P$ is well defined.

Proof of Lemma C.9. For all $t_n \to 0$, $P_n \to P$, and $H_n \to H$ in $\ell^\infty(\tilde{\mathcal{V}})$ such that $P_n \in \mathbb{D}_{\mathcal{L}}$ and $P_n + t_n H_n \in \mathbb{D}_{\mathcal{L}}$, we have that for each $(h, g) \in \bar{\mathcal{H}} \times \mathcal{G}$ with $g = (g_1, g_2)$,

$$\begin{aligned} &\mathcal{L}\left(P_n + t_n H_n\right)(h, g) - \mathcal{L}\left(P_n\right)(h, g) \\ =&\frac{t_n H_n\left(h \cdot g_2\right) P_n\left(g_2\right) - t_n P_n\left(h \cdot g_2\right) H_n\left(g_2\right)}{\left(P_n + t_n H_n\right)\left(g_2\right) P_n\left(g_2\right)} - \frac{t_n H_n\left(h \cdot g_1\right) P_n\left(g_1\right) - t_n P_n\left(h \cdot g_1\right) H_n\left(g_1\right)}{\left(P_n + t_n H_n\right)\left(g_1\right) P_n\left(g_1\right)}. \end{aligned}$$

Thus it is easy to show that

$$\lim_{n \to \infty} \sup_{(h,g) \in \bar{\mathcal{H}} \times \mathcal{G}} \left| \frac{\mathcal{L}\left(P_n + t_n H_n\right)(h, g) - \mathcal{L}\left(P_n\right)(h, g)}{t_n} - \mathcal{L}'_P(H)(h, g) \right| = 0,$$

where $\mathcal{L}'_P$ is defined as in (C.10). This implies that $\mathcal{L}$ is uniformly differentiable and verifies the derivative in (C.10). ■

Lemma C.10 *Under Assumptions 3.1 and 3.2 with $P_n, P \in \ell^\infty(\tilde{\mathcal{V}})$, we have $\sup_{v \in \tilde{\mathcal{V}}} |\sqrt{n}(P_n - P)(v) - Q_0(v)| \to 0$, where $Q_0(v) = P(vv_0)$ for all $v \in \tilde{\mathcal{V}}$ and v_0 is as in Assumption 3.2, and that $\sqrt{n}(\hat{P}_n - P)$ converges under P_n in distribution to the process $\mathbb{G}_P + Q_0$ for a tight P-Brownian bridge $\mathbb{G}_P$ with $E[\mathbb{G}_P(v_1)\mathbb{G}_P(v_2)] = P(v_1 v_2) - P(v_1)P(v_2)$ for all $v_1, v_2 \in \tilde{\mathcal{V}}$.*

Proof of Lemma C.10. The Lemma holds by Assumptions 3.1 and 3.2, the facts that $\sup_{v \in \tilde{\mathcal{V}}} |P(v)| \leq 1$ and $\sup_{v \in \tilde{\mathcal{V}}} |P_n(v^2)| \leq 1$ for all n, Lemma C.5 in this paper, and Theorem 3.10.12 of van der Vaart and Wellner (1996). ■

Lemma C.11 *Under Assumptions 3.1 and 3.2 with $P_n, P \in \ell^\infty(\tilde{\mathcal{V}})$, we have that $P_n \to P$ and that $\hat{P}_n \to P$, $\hat{\phi}_{P_n} \to \phi_P$, $T_n/n \to \Lambda(P)$, and $\hat{\sigma}_{P_n} \to \sigma_P$ almost uniformly.*

Proof of Lemma C.11. By Lemma C.10 in this paper, Hölder's inequality, and Lemma 3.10.11 of van der Vaart and Wellner (1996), we have that

$$\begin{aligned} \|P_n - P\|_\infty \leq& \|P_n - P - n^{-1/2} Q_0\|_\infty + \|n^{-1/2} Q_0\|_\infty \\ \leq& \|P_n - P - n^{-1/2} Q_0\|_\infty + n^{-1/2} \sup_{v \in \tilde{\mathcal{V}}} |P(v^2) P(v_0^2)|^{1/2} \to 0, \end{aligned}$$

where Q_0 is the function defined in Lemma C.10. By Lemma C.6 in this paper and Lemma 1.9.3 of van der Vaart and Wellner (1996), $\|\hat{P}_n - P_n\|_\infty \to 0$ almost uniformly. Then we have that $\|\hat{P}_n - P\|_\infty \to 0$ almost uniformly. The rest of the results follow from the constructions of $\hat{\phi}_{P_n}$, T_n/n, and $\hat{\sigma}_{P_n}$. By the construction of $\bar{\mathcal{H}}$, the $\sigma_Q^2(h, g)$ in (19) can also be written as

$$\sigma_Q^2(h, g) = \Lambda(Q) \cdot \left\{ \frac{|Q\left(h \cdot g_2\right)|}{Q^2\left(g_2\right)} - \frac{Q^2\left(h \cdot g_2\right)}{Q^3\left(g_2\right)} + \frac{|Q\left(h \cdot g_1\right)|}{Q^2\left(g_1\right)} - \frac{Q^2\left(h \cdot g_1\right)}{Q^3\left(g_1\right)} \right\}. \quad \text{(C.11)}$$

Similar to (C.11), we can write the $\hat{\sigma}^2_{P_n}(h,g)$ in (21) as

$$\hat{\sigma}^2_{P_n}(h,g) = \frac{T_n}{n} \cdot \left\{ \frac{|\hat{P}_n(h \cdot g_2)|}{\hat{P}^2_n(g_2)} - \frac{\hat{P}^2_n(h \cdot g_2)}{\hat{P}^3_n(g_2)} + \frac{|\hat{P}_n(h \cdot g_1)|}{\hat{P}^2_n(g_1)} - \frac{\hat{P}^2_n(h \cdot g_1)}{\hat{P}^3_n(g_1)} \right\}. \tag{C.12}$$

Then the almost uniform convergence of $\hat{P}_n$ to P in $\ell^\infty(\tilde{\mathcal{V}})$ implies the almost uniform convergence of the $\hat{\sigma}^2_{P_n}$ in (C.12) to the σ^2_P as in (C.11). ■

Proof of Lemma 3.1. By the Hadamard derivative of $\mathcal{L}$ in (C.10), together with Lemma C.10 in this paper and Theorem 3.9.4 (delta method) of van der Vaart and Wellner (1996), we have that under P_n,

$$\sqrt{n}(\hat{\phi}_{P_n} - \phi_P) = \sqrt{n}\{\mathcal{L}(\hat{P}_n) - \mathcal{L}(P)\} \rightsquigarrow \mathcal{L}'_P(\mathbb{G}_P + Q_0). \tag{C.13}$$

By Lemma C.11, $T_n/n \to \Lambda(P)$ almost uniformly. Thus by Lemmas 1.9.3(ii) and 1.10.2(iii), Example 1.4.7 (Slutsky's lemma), and Theorem 1.3.6 (continuous mapping) of van der Vaart and Wellner (1996),

$$\sqrt{T_n}(\hat{\phi}_{P_n} - \phi_P) = \sqrt{T_n/n} \cdot \sqrt{n}(\hat{\phi}_{P_n} - \phi_P) \rightsquigarrow \Lambda(P)^{1/2}\mathcal{L}'_P(\mathbb{G}_P + Q_0). \tag{C.14}$$

Let $\mathbb{G} = \Lambda(P)^{1/2}\mathcal{L}'_P(\mathbb{G}_P + Q_0)$. Then $\mathbb{G}$ is tight, because $\mathbb{G}_P$ is tight and $\mathcal{L}'_P$ is a continuous map. Thus (C.14) verifies the first claim of Lemma 3.1. Now we show the continuity of $\mathbb{G}$ under ρ_P. Define a semimetric on $\tilde{\mathcal{V}}$ by

$$\rho_2(v,v') = E\left[|\mathbb{G}_P(v) - \mathbb{G}_P(v')|^2\right]^{1/2}$$

for all $v, v' \in \tilde{\mathcal{V}}$. This semimetric is the one defined in van der Vaart and Wellner (1996, p. 39) with $p = 2$. Since $\mathbb{G}_P$ is tight, it follows from the discussion in Example 1.5.10 of van der Vaart and Wellner (1996) that $\mathbb{G}_P$ almost surely has a uniformly ρ_2-continuous path. Since $\mathbb{G}_P$ is a P-Brownian bridge,

$$\rho_2^2(v,v') = P((v-v')^2) - P^2(v-v') \leq \|v-v'\|^2_{L^2(P)} \tag{C.15}$$

for all $v, v' \in \tilde{\mathcal{V}}$. Therefore, $\mathbb{G}_P$ almost surely has a uniformly continuous path under $\|\cdot\|_{L^2(P)}$. By Lemma 3.10.11 of van der Vaart and Wellner (1996), $P(v_0) = 0$ and $P(v_0^2) < \infty$, where v_0 is as in Assumption 3.2. Hölder's inequality implies that for every $v \in L^2(P)$, $\|v \cdot 1\|_{L^1(P)} \leq 1 \cdot \|v\|_{L^2(P)}$. By Hölder's inequality, P and Q_0 are both continuous on $\tilde{\mathcal{V}}$ under $\|\cdot\|_{L^2(P)}$, where Q_0 is as in Lemma C.10. Suppose that there are $(h,g), (h',g') \in \bar{\mathcal{H}} \times \mathcal{G}$ with

$g = (g_1, g_2)$ and $g' = (g'_1, g'_2)$. Then for $j \in \{1, 2\}$ we have

$$\begin{aligned} &\left\|g_j - g'_j\right\|_{L^2(P)} \leq \rho_P\left((h, g), (h', g')\right) \text{ and} \\ &\left\|h \cdot g_j - h' \cdot g'_j\right\|_{L^2(P)} \leq \left\|h - h'\right\|_{L^2(P)} + \left\|g_j - g'_j\right\|_{L^2(P)} \leq \rho_P\left((h, g), (h', g')\right). \end{aligned} \tag{C.16}$$

By (C.10) and (C.16), together with the continuity of $\mathbb{G}_P$, P, and Q_0 under $\|\cdot\|_{L^2(P)}$, we conclude that $\mathbb{G}$ almost surely has a continuous path under ρ_P.

Next, we show the variance of $\mathbb{G}(h, g)$ for each $(h, g) \in \bar{\mathcal{H}} \times \mathcal{G}$ with $g = (g_1, g_2)$. Since $\mathcal{L}'_P(H)$ is linear in H, $Var(\mathbb{G}(h, g)) = \Lambda(P) \cdot Var(\mathcal{L}'_P(\mathbb{G}_P)(h, g))$. First, we have that

$$\begin{aligned} &Var(\mathcal{L}'_P(\mathbb{G}_P)(h, g)) \\ &= E\left[\left(\frac{\mathbb{G}_P(h \cdot g_2) P(g_2) - P(h \cdot g_2) \mathbb{G}_P(g_2)}{P^2(g_2)} - \frac{\mathbb{G}_P(h \cdot g_1) P(g_1) - P(h \cdot g_1) \mathbb{G}_P(g_1)}{P^2(g_1)}\right)^2\right]. \end{aligned} \tag{C.17}$$

Since $\mathbb{G}_P$ is a Brownian bridge with $E[\mathbb{G}_P(v_1)\mathbb{G}_P(v_2)] = P(v_1 v_2) - P(v_1)P(v_2)$ for all $v_1, v_2 \in \tilde{\mathcal{V}}$, we have

$$\begin{aligned} &E\left[\left(\frac{\mathbb{G}_P(h \cdot g_2) P(g_2) - P(h \cdot g_2) \mathbb{G}_P(g_2)}{P^2(g_2)}\right)^2\right] \\ &= \frac{P(h^2 \cdot g_2) - P^2(h \cdot g_2)}{P^2(g_2)} + \frac{P^2(h \cdot g_2)}{P^3(g_2)} - \frac{P^2(h \cdot g_2)}{P^2(g_2)} - \frac{2P^2(h \cdot g_2)}{P^3(g_2)} + \frac{2P^2(h \cdot g_2)}{P^2(g_2)} \\ &= \frac{P(h^2 \cdot g_2)}{P^2(g_2)} - \frac{P^2(h \cdot g_2)}{P^3(g_2)}. \end{aligned} \tag{C.18}$$

Similarly,

$$E\left[\left(\frac{\mathbb{G}_P(h \cdot g_1) P(g_1) - P(h \cdot g_1) \mathbb{G}_P(g_1)}{P^2(g_1)}\right)^2\right] = \frac{P(h^2 \cdot g_1)}{P^2(g_1)} - \frac{P^2(h \cdot g_1)}{P^3(g_1)}. \tag{C.19}$$

Also, we have that

$$\begin{aligned} &E\left[\left(\mathbb{G}_P(h \cdot g_2) P(g_2) - P(h \cdot g_2) \mathbb{G}_P(g_2)\right)\left(\mathbb{G}_P(h \cdot g_1) P(g_1) - P(h \cdot g_1) \mathbb{G}_P(g_1)\right)\right] \\ &= P(g_2) P(g_1) P(h^2 g_2 g_1) - P(g_2) P(h g_1) P(h g_2 g_1) - P(h g_2) P(g_1) P(h g_2 g_1) \\ &\quad + P(h g_2) P(h g_1) P(g_2 g_1) = 0, \end{aligned} \tag{C.20}$$

where we use the fact that $g_1 g_2 = 0$ by the construction of $\mathcal{G}$. By (C.20), the expectation on the right-hand side of (C.17) is equal to the sum of the expectations in (C.18) and (C.19).

Thus we now have that

$$Var(\mathcal{L}'_P(\mathbb{G}_P)(h,g)) = \frac{P\left(h^2 \cdot g_2\right)}{P^2(g_2)} - \frac{P^2(h \cdot g_2)}{P^3(g_2)} + \frac{P\left(h^2 \cdot g_1\right)}{P^2(g_1)} - \frac{P^2(h \cdot g_1)}{P^3(g_1)},$$

which, together with $Var(\mathbb{G}(h,g)) = \Lambda(P) \cdot Var(\mathcal{L}'_P(\mathbb{G}_P)(h,g))$, verifies the equality that $Var(\mathbb{G}(h,g)) = \sigma_P^2(h,g)$ for the σ_P^2 in (19). For every $(h,g) \in \bar{\mathcal{H}} \times \mathcal{G}$ with $g = (g_1, g_2)$,

$$\begin{aligned}\sigma_P^2(h,g) &= \Lambda(P)\left\{\frac{P\left(h^2 \cdot g_2\right)}{P^2(g_2)} - \frac{P^2(h \cdot g_2)}{P^3(g_2)} + \frac{P\left(h^2 \cdot g_1\right)}{P^2(g_1)} - \frac{P^2(h \cdot g_1)}{P^3(g_1)}\right\} \\ &= \frac{\Lambda(P)}{P(g_2)}\frac{|P(h \cdot g_2)|}{P(g_2)}\left[1 - \frac{|P(h \cdot g_2)|}{P(g_2)}\right] + \frac{\Lambda(P)}{P(g_1)}\frac{|P(h \cdot g_1)|}{P(g_1)}\left[1 - \frac{|P(h \cdot g_1)|}{P(g_1)}\right].\end{aligned}$$

Then $\sigma_P^2(h,g) \leq 1/4 \cdot \{\Lambda(P)/P(g_2) + \Lambda(P)/P(g_1)\}$, since $0 \leq |P(hg_j)|/P(g_j) \leq 1$ for $j \in \{1,2\}$. Recall that K is the number of elements in $\mathcal{Z}$. We have that for each $j \in \{1,2\}$,

$$\frac{\Lambda(P)}{P(g_j)} \leq \max_{1 \leq k' \leq K} \frac{\prod_{k=1}^{K} P\left(1_{\mathbb{R}\times\mathbb{R}\times\{z_k\}}\right)}{P\left(1_{\mathbb{R}\times\mathbb{R}\times\{z_{k'}\}}\right)} \leq \left(\frac{1}{K-1}\right)^{K-1},$$

which implies that

$$\sigma_P^2(h,g) \leq 1/4 \cdot \max_{(g_1',g_2') \in \mathcal{G}} \left\{\Lambda(P)/P\left(g_2'\right) + \Lambda(P)/P\left(g_1'\right)\right\} \leq 1/2 \cdot (K-1)^{-(K-1)}.$$

When $K = 2$, $\sigma_P^2(h,g) \leq 1/4$ by the construction of $\Lambda(P)$. ■

Lemma C.12 *Under ρ_P, ϕ_P and σ_P are continuous on $\bar{\mathcal{H}} \times \mathcal{G}$.*

Proof of Lemma C.12. Suppose there are $(h,g), (h^k, g^k) \in \bar{\mathcal{H}} \times \mathcal{G}$ with $g = (g_1, g_2)$, $g^k = (g_1^k, g_2^k)$, and $(h^k, g^k) \to (h,g)$ under ρ_P. Since $\mathcal{G}_K$ is finite, $(h^k, g^k) \to (h,g)$ under ρ_P implies that $P(g_j^k) = P(g_j)$ for each $j \in \{1,2\}$ when k is sufficiently large. If $P(g_j) = 0$,[5] then by (13) $P(h \cdot g_j)/P(g_j) = 0$, $P(h^k \cdot g_j^k)/P(g_j^k) = 0$ when k is large, and

$$\left|\frac{P(h \cdot g_j)}{P(g_j)} - \frac{P(h^k \cdot g_j^k)}{P(g_j^k)}\right| = 0.$$

If $P(g_j) \neq 0$, then for each $j \in \{1,2\}$ and large k, $P(g_j^k) = P(g_j) \neq 0$ and

$$\left|\frac{P(h \cdot g_j)}{P(g_j)} - \frac{P(h^k \cdot g_j^k)}{P(g_j^k)}\right| \leq \frac{\left\|h \cdot g_j - h^k \cdot g_j^k\right\|_{L^2(P)}}{P(g_j)} \leq \frac{\rho_P((h,g),(h^k,g^k))}{P(g_j)}$$

[5] If $P(g_j) = 0$ for some $g_j \in \mathcal{G}_K$, then $\Lambda(P) = 0$, which is a trivial case. We consider this case only for the sake of completeness.

by Hölder's inequality and (C.16). Thus we can conclude that

$$\left|\phi_P(h,g) - \phi_P(h^k,g^k)\right| = \left|\left(\frac{P(h\cdot g_2)}{P(g_2)} - \frac{P(h\cdot g_1)}{P(g_1)}\right) - \left(\frac{P(h^k\cdot g_2^k)}{P(g_2^k)} - \frac{P(h^k\cdot g_1^k)}{P(g_1^k)}\right)\right| \to 0$$

if $(h^k,g^k)\to(h,g)$ under ρ_P. Similarly, we can show that σ_P is continuous on $\bar{\mathcal{H}}\times\mathcal{G}$ under ρ_P. ∎

We define some new notation which will be used in the following results. Define a random element $\hat{\varphi}_P:\Omega\to\ell^\infty\left(\Xi\times\bar{\mathcal{H}}\times\mathcal{G}\right)$ such that for each $\omega\in\Omega$ and each $(\xi,h,g)\in\Xi\times\bar{\mathcal{H}}\times\mathcal{G}$,

$$\hat{\varphi}_P(\omega)(\xi,h,g) = \frac{\phi_P(h,g)}{\mathcal{M}\left(\hat{\sigma}_{P_n}(\omega)\right)(\xi,h,g)}, \tag{C.21}$$

and let $\varphi_P\in\ell^\infty\left(\Xi\times\bar{\mathcal{H}}\times\mathcal{G}\right)$ be such that for each $(\xi,h,g)\in\Xi\times\bar{\mathcal{H}}\times\mathcal{G}$,

$$\varphi_P(\xi,h,g) = \frac{\phi_P(h,g)}{\mathcal{M}\left(\sigma_P\right)(\xi,h,g)}.$$

Here, $\hat{\sigma}_{P_n}$ is estimated from data, hence it depends on ω, and so does $\hat{\varphi}_P$. When there is no danger of confusion, we omit the ω from $\hat{\sigma}_{P_n}$ and $\hat{\varphi}_P$ for brevity. Given each sequence $r_n\to\infty$ and each ν which satisfies Assumption 3.3, define

$$\mathbb{D}_n(\omega) = \left\{\psi\in\ell^\infty\left(\Xi\times\bar{\mathcal{H}}\times\mathcal{G}\right) : \mathcal{S}\left(\hat{\varphi}_P(\omega)+r_n^{-1}\psi\right)\in L^1(\nu)\right\} \tag{C.22}$$

for all $\omega\in\Omega$, and

$$g_n(\omega)(\psi) = r_n\mathcal{I}\circ\mathcal{S}\left(\hat{\varphi}_P(\omega)+r_n^{-1}\psi\right) \tag{C.23}$$

for all $\omega\in\Omega$ and all $\psi\in\mathbb{D}_n(\omega)$. Here, g_n also depends on ω; for brevity, we omit ω from g_n as well. If the H_0 in (15) is true with $Q=P_n$ for all n, then $\mathcal{S}(\hat{\varphi}_P)=0$ by Lemma C.13, and so $g_n(\psi)=r_n\left\{\mathcal{I}\circ\mathcal{S}\left(\hat{\varphi}_P+r_n^{-1}\psi\right)-\mathcal{I}\circ\mathcal{S}\left(\hat{\varphi}_P\right)\right\}$. Define a correspondence $\Psi:\Xi\times\ell^\infty\left(\Xi\times\bar{\mathcal{H}}\times\mathcal{G}\right)\twoheadrightarrow\bar{\mathcal{H}}\times\mathcal{G}$ by

$$\Psi(\xi,\psi) = \left\{(h,g)\in\bar{\mathcal{H}}\times\mathcal{G} : \psi(\xi,h,g) = \mathcal{S}(\psi)(\xi)\right\} \tag{C.24}$$

for all $\xi\in\Xi$ and all $\psi\in\ell^\infty\left(\Xi\times\bar{\mathcal{H}}\times\mathcal{G}\right)$, and define a metric $\rho_{\xi\psi}$ on $\Xi\times\ell^\infty\left(\Xi\times\bar{\mathcal{H}}\times\mathcal{G}\right)$ by

$$\rho_{\xi\psi}((\xi_1,\psi_1),(\xi_2,\psi_2)) = |\xi_1-\xi_2| + \|\psi_1-\psi_2\|_\infty \tag{C.25}$$

for all $(\xi_1, \psi_1), (\xi_2, \psi_2) \in \Xi \times \ell^\infty\left(\Xi \times \bar{\mathcal{H}} \times \mathcal{G}\right)$. Also, define a metric on $\Xi \times \bar{\mathcal{H}} \times \mathcal{G}$ by

$$\rho_{\xi hg}((\xi_1, h_1, g_1), (\xi_2, h_2, g_2)) = |\xi_1 - \xi_2| + \rho_P((h_1, g_1), (h_2, g_2)) \tag{C.26}$$

for all $(\xi_1, h_1, g_1), (\xi_2, h_2, g_2) \in \Xi \times \bar{\mathcal{H}} \times \mathcal{G}$. For every set $A \subset \bar{\mathcal{H}} \times \mathcal{G}$ and every $\delta > 0$, define

$$A^\delta = \left\{(h, g) \in \bar{\mathcal{H}} \times \mathcal{G} : \inf_{(h', g') \in A} \rho_P\left((h, g), (h', g')\right) \leq \delta\right\}. \tag{C.27}$$

Lemma C.13 *Suppose Assumption 3.2 holds and the H_0 in (15) is true with $Q = P_n$ for all n. Then the H_0 in (15) is true with $Q = P$. This implies that $\sup_{(h,g)\in\bar{\mathcal{H}}\times\mathcal{G}} \phi_P(h, g) = 0$, and hence that $\mathcal{S}(\varphi_P) = 0$ and $\mathcal{S}(\hat{\varphi}_P) = 0$ for all $\omega \in \Omega$.*

Proof of Lemma C.13. By Lemma C.11, we have $\|P_n - P\|_\infty \to 0$. Thus $\phi_{P_n} \to \phi_P$ in $\ell^\infty(\bar{\mathcal{H}} \times \mathcal{G})$, and by the assumption that $\sup_{(h,g)\in\mathcal{H}\times\mathcal{G}} \phi_{P_n}(h, g) \leq 0$ for all n, we have that $\sup_{(h,g)\in\mathcal{H}\times\mathcal{G}} \phi_P(h, g) \leq 0$. This implies that $\sup_{(h,g)\in\bar{\mathcal{H}}\times\mathcal{G}} \phi_P(h, g) \leq 0$ by the constructions of ϕ_P and $\bar{\mathcal{H}}$. By the construction of $\bar{\mathcal{H}} \times \mathcal{G}$, there is some $(h, g) \in \bar{\mathcal{H}} \times \mathcal{G}$, such as $h = 1_{\{a\}\times\{0\}\times\mathbb{R}}$ for some $a \in \mathbb{R}$, for which $\phi_P(h, g) = 0$. Therefore, $\sup_{(h,g)\in\bar{\mathcal{H}}\times\mathcal{G}} \phi_P(h, g) = 0$ under the assumptions. Because $\xi \in \Xi$ is always positive by the construction of Ξ, we have that $\mathcal{S}(\varphi_P)(\xi) = 0$ for all $\xi \in \Xi$. For the same reason, $\mathcal{S}(\hat{\varphi}_P)(\xi) = 0$ for all $\xi \in \Xi$ and all $\omega \in \Omega$. ∎

Lemma C.14 *The correspondence Ψ defined in (C.24) is upper hemicontinuous[6] at (ξ, φ_P) for all $\xi \in \Xi$. In addition, suppose the H_0 in (15) is true with $Q = P$. Then for every $\delta > 0$ there is an $\varepsilon > 0$ such that $\Psi(\xi', \psi) \subset \Psi(\xi, \varphi_P)^\delta$ (where the latter is defined as in (C.27)) for all $\xi, \xi' \in \Xi$ and all $\psi \in \ell^\infty\left(\Xi \times \bar{\mathcal{H}} \times \mathcal{G}\right)$ with $\|\psi - \varphi_P\|_\infty < \varepsilon$.*

Proof of Lemma C.14. We first show that Ψ is upper hemicontinuous at (ξ, φ_P) for all $\xi \in \Xi$. We do this **in three steps**. **First**, we show that $\Psi(\xi, \varphi_P)$ is compact for each $\xi \in \Xi$ under ρ_P. Clearly, given an arbitrary $\xi \in \Xi$, $\varphi_P(\xi, \cdot, \cdot)$ is continuous on $\bar{\mathcal{H}} \times \mathcal{G}$ under ρ_P by Lemma C.12. Because $\bar{\mathcal{H}} \times \mathcal{G}$ is compact by Lemma C.8, $\Psi(\xi, \varphi_P)$ is not empty. Since $\Psi(\xi, \varphi_P) \subset \bar{\mathcal{H}} \times \mathcal{G}$, it suffices to show that $\Psi(\xi, \varphi_P)$ is closed in $\bar{\mathcal{H}} \times \mathcal{G}$. Fix $\xi \in \Xi$. Suppose there is a sequence $\{(h_n, g_n)\}_n \subset \Psi(\xi, \varphi_P)$ such that $(h_n, g_n) \to (h, g) \in \bar{\mathcal{H}} \times \mathcal{G}$ under ρ_P. Then for all n, $\varphi_P(\xi, h_n, g_n) = \mathcal{S}(\varphi_P)(\xi)$. Since $\varphi_P(\xi, \cdot, \cdot)$ is continuous by Lemma C.12, $\varphi_P(\xi, h_n, g_n) \to \varphi_P(\xi, h, g)$ as $(h_n, g_n) \to (h, g)$. Thus $\varphi_P(\xi, h, g) = \mathcal{S}(\varphi_P)(\xi)$, which implies that $\Psi(\xi, \varphi_P)$ is closed in $\bar{\mathcal{H}} \times \mathcal{G}$ and therefore compact. **Second**, we show that if there is a sequence $\{(\xi_n, \psi_n), (h_n, g_n)\}$ such that $(h_n, g_n) \in \Psi(\xi_n, \psi_n)$ and $\rho_{\xi\psi}((\xi_n, \psi_n), (\xi, \varphi_P)) \to 0$, where $\rho_{\xi\psi}$ is defined in (C.25), then (h_n, g_n) has a limit point[7]

[6] See Definition 17.2 of upper hemicontinuity in Aliprantis and Border (2006).

[7] See the definition of limit point in Aliprantis and Border (2006, p. 31).

in $\Psi(\xi, \varphi_P)$. Notice that by the constructions of Ξ and $\bar{\mathcal{H}} \times \mathcal{G}$, $\Xi \times \bar{\mathcal{H}} \times \mathcal{G}$ is compact under the metric $\rho_{\xi hg}$ defined in (C.26). It is easy to show, by Lemma C.12, that φ_P is continuous on $\Xi \times \bar{\mathcal{H}} \times \mathcal{G}$ under $\rho_{\xi hg}$, and hence that it is uniformly continuous. Thus $\rho_{\xi\psi}((\xi_n, \psi_n), (\xi, \varphi_P)) \to 0$ implies that

$$\begin{aligned}|\mathcal{S}(\psi_n)(\xi_n) - \mathcal{S}(\varphi_P)(\xi)| \leq & \sup_{(h,g)\in\bar{\mathcal{H}}\times\mathcal{G}} |\psi_n(\xi_n, h, g) - \varphi_P(\xi_n, h, g)| \\ & + \sup_{(h,g)\in\bar{\mathcal{H}}\times\mathcal{G}} |\varphi_P(\xi_n, h, g) - \varphi_P(\xi, h, g)| \to 0,\end{aligned}$$

where $\sup_{(h,g)\in\bar{\mathcal{H}}\times\mathcal{G}} |\varphi_P(\xi_n, h, g) - \varphi_P(\xi, h, g)|$ converges to 0 because φ_P is uniformly continuous on $\Xi \times \bar{\mathcal{H}} \times \mathcal{G}$ under $\rho_{\xi hg}$. This implies that $\psi_n(\xi_n, h_n, g_n) \to \mathcal{S}(\varphi_P)(\xi)$. Suppose, by way of contradiction, that (h_n, g_n) has no limit point in $\Psi(\xi, \varphi_P)$. This implies that for each $(h, g) \in \Psi(\xi, \varphi_P)$ there exists an open neighborhood $V_{h,g}$ of (h, g) and an $n_{h,g}$ such that $(h_n, g_n) \notin V_{h,g}$ when $n \geq n_{h,g}$. Because we have shown that $\Psi(\xi, \varphi_P)$ is compact in $\bar{\mathcal{H}} \times \mathcal{G}$, there is a finite open cover V such that $\Psi(\xi, \varphi_P) \subset V = V_{h^1,g^1} \cup \cdots \cup V_{h^M,g^M}$. Let $n_0 = \max_{m\leq M} n_{h^m,g^m}$. Thus if $n > n_0$, then $(h_n, g_n) \notin V$, and hence $(h_n, g_n) \notin \Psi(\xi, \varphi_P)$. Since $\bar{\mathcal{H}} \times \mathcal{G}$ is compact and V^c is closed in $\bar{\mathcal{H}} \times \mathcal{G}$, V^c is compact. Notice that $V^c \cap \Psi(\xi, \varphi_P) = \varnothing$. Thus

$$\sup_{(h,g)\in V^c} \varphi_P(\xi, h, g) < \sup_{(h,g)\in\bar{\mathcal{H}}\times\mathcal{G}} \varphi_P(\xi, h, g) = \sup_{(h,g)\in\Psi(\xi,\varphi_P)} \varphi_P(\xi, h, g).$$

Let $\delta = \sup_{(h,g)\in\bar{\mathcal{H}}\times\mathcal{G}} \varphi_P(\xi, h, g) - \sup_{(h,g)\in V^c} \varphi_P(\xi, h, g)$. Recall that $(h_n, g_n) \in V^c$ for all $n > n_0$. Thus $\psi_n(\xi_n, h_n, g_n) = \sup_{(h,g)\in\bar{\mathcal{H}}\times\mathcal{G}} \psi_n(\xi_n, h, g) = \sup_{(h,g)\in V^c} \psi_n(\xi_n, h, g)$, so

$$\begin{aligned}\left|\psi_n(\xi_n, h_n, g_n) - \sup_{(h,g)\in V^c} \varphi_P(\xi, h, g)\right| \leq & \sup_{(h,g)\in\bar{\mathcal{H}}\times\mathcal{G}} |\psi_n(\xi_n, h, g) - \varphi_P(\xi_n, h, g)| \\ & + \sup_{(h,g)\in\bar{\mathcal{H}}\times\mathcal{G}} |\varphi_P(\xi_n, h, g) - \varphi_P(\xi, h, g)| \to 0.\end{aligned}$$

This implies that for sufficiently large n,

$$\psi_n(\xi_n, h_n, g_n) \leq \sup_{(h,g)\in V^c} \varphi_P(\xi, h, g) + \frac{\delta}{2} = \sup_{(h,g)\in\bar{\mathcal{H}}\times\mathcal{G}} \varphi_P(\xi, h, g) - \frac{\delta}{2}.$$

This contradicts $\psi_n(\xi_n, h_n, g_n) \to \mathcal{S}(\varphi_P)(\xi)$. Thus (h_n, g_n) has a limit point in $\Psi(\xi, \varphi_P)$. **Third**, by Theorem 17.20(ii) of Aliprantis and Border (2006), together with the fact that $\Xi \times \ell^\infty\left(\Xi \times \bar{\mathcal{H}} \times \mathcal{G}\right)$ is first countable under the metric $\rho_{\xi\psi}$ defined in (C.25) (every metric space is first countable), Ψ is upper hemicontinuous at (ξ, φ_P).

Now we prove the second claim in the Lemma. Fix $\delta > 0$. Since Ψ is upper hemicontinuous at (ξ, φ_P) for all $\xi \in \Xi$, we have that for each ξ there is an open ball $B_{\varepsilon_\xi}(\xi, \varphi_P)$

under $\rho_{\xi\psi}$ with center (ξ, φ_P) and radius ε_ξ such that $\Psi(\xi', \varphi') \subset \Psi(\xi, \varphi_P)^\delta$ for all $(\xi', \varphi') \in B_{\varepsilon_\xi}(\xi, \varphi_P)$, where $\Psi(\xi, \varphi_P)^\delta$ is defined as in (C.27). Notice that $\{B_{\varepsilon_\xi/2}(\xi)\}_{\xi\in\Xi}$ is an open cover of Ξ, where each $B_{\varepsilon_\xi/2}(\xi)$ is an open ball in $\mathbb{R}$ with center ξ and radius $\varepsilon_\xi/2$. Since Ξ is compact by construction, there is a finite open cover $\{B_{\varepsilon_i}(\xi_i)\}_{i=1}^M$ of Ξ with $\varepsilon_i = \varepsilon_{\xi_i}/2$. Let $\varepsilon = \min_{i\leq M}\varepsilon_i$. Then for every $\xi' \in \Xi$ and every $\psi \in \ell^\infty\left(\Xi \times \bar{\mathcal{H}} \times \mathcal{G}\right)$ with $\|\psi - \varphi_P\|_\infty < \varepsilon$, there is an open ball $B_{\varepsilon_{\xi_i}}(\xi_i, \varphi_P)$ such that $(\xi', \psi) \subset B_{\varepsilon_{\xi_i}}(\xi_i, \varphi_P)$. This implies that $\Psi(\xi', \psi) \subset \Psi(\xi_i, \varphi_P)^\delta$. Suppose the H_0 in (15) is true with $Q = P$. By Lemma C.13, we have that $\mathcal{S}(\varphi_P) = 0$ and

$$\Psi(\xi, \varphi_P) = \Psi(\tilde{\xi}, \varphi_P) = \left\{(h, g) \in \bar{\mathcal{H}} \times \mathcal{G} : \phi_P(h, g) = 0\right\}$$

for all $\xi, \tilde{\xi} \in \Xi$. Thus $\Psi(\xi', \psi) \subset \Psi(\xi, \varphi_P)^\delta$ for all $\xi \in \Xi$, that is, the second claim holds. ■

Lemma C.15 *Suppose Assumptions 3.1, 3.2, and 3.3 hold and the H_0 in (15) is true with $Q = P_n$ for all n. For every $\varepsilon > 0$, there is a measurable set $\Omega_0 \subset \Omega$ with $\mathbb{P}(\Omega_0) \geq 1 - \varepsilon$ such that for every subsequence $\{\psi_{n_m}\}$ with $\psi_{n_m} \in \mathbb{D}_{n_m}(\omega_{n_m})$, $\omega_{n_m} \in \Omega_0$, where $\mathbb{D}_{n_m}(\omega_{n_m})$ is defined in (C.22), and $\psi_{n_m} \to \psi$ for some $\psi \in C\left(\Xi \times \bar{\mathcal{H}} \times \mathcal{G}\right)$ under the $\rho_{\xi hg}$ defined in (C.26), we have that*

$$g_{n_m}(\omega_{n_m})(\psi_{n_m}) \to \mathcal{I} \circ \mathcal{S}_{\Psi(\xi,\varphi_P)}(\psi),$$

where g_{n_m} is defined in (C.23).[8]

Proof of Lemma C.15. For simplicity of notation, we replace n_m with n. Note that all the following results hold for every subsequence indexed by n_m. By Lemma C.8, $\bar{\mathcal{H}} \times \mathcal{G}$ is compact under ρ_P. By Lemma C.11, we have $\hat{\sigma}_{P_n} \to \sigma_P$ almost uniformly. Then by construction, $\hat{\varphi}_P \to \varphi_P$ almost uniformly, where $\hat{\varphi}_P$ is defined in (C.21). By Lemma C.13, $\mathcal{S}(\varphi_P) = 0$ and $\mathcal{S}(\hat{\varphi}_P) = 0$ for all $\omega \in \Omega$. For every $\psi \in C\left(\Xi \times \bar{\mathcal{H}} \times \mathcal{G}\right)$, since $\hat{\varphi}_P(\xi, \cdot, \cdot) + r_n^{-1}\psi(\xi, \cdot, \cdot)$ may not be continuous on $\bar{\mathcal{H}} \times \mathcal{G}$, $\Psi\left(\xi, \hat{\varphi}_P + r_n^{-1}\psi\right)$ may be empty. Here, we construct a modified version of $\hat{\varphi}_P$, denoted by $\tilde{\varphi}_P$, such that

(i) $\tilde{\varphi}_P(\xi, \cdot, \cdot)$ is upper semicontinuous for every $\omega \in \Omega$, every n, and every $\xi \in \Xi$;

(ii) $\sup_{(h,g)\in\bar{\mathcal{H}}\times\mathcal{G}} \hat{\varphi}_P(\xi, h, g) = \sup_{(h,g)\in\bar{\mathcal{H}}\times\mathcal{G}} \tilde{\varphi}_P(\xi, h, g)$ for every $\omega \in \Omega$, every n, and every $\xi \in \Xi$;

(iii) $\sup_{(h,g)\in\bar{\mathcal{H}}\times\mathcal{G}} \left(\hat{\varphi}_P + r_n^{-1}\psi\right)(\xi, h, g) = \sup_{(h,g)\in\bar{\mathcal{H}}\times\mathcal{G}} \left(\tilde{\varphi}_P + r_n^{-1}\psi\right)(\xi, h, g)$ for every function $\psi \in C\left(\Xi \times \bar{\mathcal{H}} \times \mathcal{G}\right)$, every $\omega \in \Omega$, every n, and every $\xi \in \Xi$;

[8]Lemma C.15 implies that under H_0, $\mathcal{I} \circ \mathcal{S}_{\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}}$ is the Hadamard directional derivative of $\mathcal{I} \circ \mathcal{S}$ at φ_P. See the definition of Hadamard directional differentiability in Shapiro (1990).

(iv) for every $\varepsilon > 0$ there is a measurable set $A \subset \Omega$ with $\mathbb{P}(A) \geq 1 - \varepsilon$ such that for all $\varphi \in \ell^\infty\left(\Xi \times \bar{\mathcal{H}} \times \mathcal{G}\right)$, $\tilde{\varphi}_P + r_n^{-1}\varphi \to \varphi_P$ uniformly on A.

Specifically, for all $\omega \in \Omega$, all $(\xi, h, g) \in \Xi \times \bar{\mathcal{H}} \times \mathcal{G}$, and all n, we define $\tilde{\varphi}_P(\xi, h, g)$ by

$$\tilde{\varphi}_P(\xi, h, g) = \lim_{\delta \downarrow 0} \sup_{(h',g') \in B_\delta(h,g)} \hat{\varphi}_P(\xi, h', g'), \tag{C.28}$$

where $B_\delta(h, g)$ is an open ball in $\bar{\mathcal{H}} \times \mathcal{G}$ under ρ_P with center (h, g) and radius δ.

Fix $\omega \in \Omega$, n, and $\xi \in \Xi$. **First**, we prove (i), that is, $\tilde{\varphi}_P(\xi, \cdot, \cdot)$ is upper semicontinuous at every $(h, g) \in \bar{\mathcal{H}} \times \mathcal{G}$. Fix $(h, g) \in \bar{\mathcal{H}} \times \mathcal{G}$. By (C.28), for each $\varepsilon > 0$, there is a $\delta_\varepsilon > 0$ such that

$$\hat{\varphi}_P\left(\xi, h', g'\right) \leq \tilde{\varphi}_P(\xi, h, g) + \frac{\varepsilon}{2} \tag{C.29}$$

for all $(h', g') \in B_{\delta_\varepsilon}(h, g)$, where $B_{\delta_\varepsilon}(h, g)$ denotes the open ball in $\bar{\mathcal{H}} \times \mathcal{G}$ under ρ_P with center (h, g) and radius δ_ε. Fix $(h_1, g_1) \in B_{\delta_\varepsilon/2}(h, g)$. By definition, there is a $\delta_2 > 0$ such that for all δ' with $0 < \delta' \leq \delta_2$,

$$\tilde{\varphi}_P(\xi, h_1, g_1) \leq \sup_{(h_2,g_2) \in B_{\delta'}(h_1,g_1)} \hat{\varphi}_P(\xi, h_2, g_2) + \frac{\varepsilon}{2}.$$

Let $\delta = \min\{\delta_\varepsilon/2, \delta_2\}$. Then for this (h_1, g_1), we have that

$$\tilde{\varphi}_P(\xi, h_1, g_1) \leq \sup_{(h_2,g_2) \in B_{\delta}(h_1,g_1)} \hat{\varphi}_P(\xi, h_2, g_2) + \frac{\varepsilon}{2}.$$

Notice that if $(h_2, g_2) \in B_\delta(h_1, g_1)$, then $(h_2, g_2) \in B_{\delta_\varepsilon}(h, g)$, and hence $\hat{\varphi}_P(\xi, h_2, g_2) \leq \tilde{\varphi}_P(\xi, h, g) + \varepsilon/2$. This implies that $\sup_{(h_2,g_2) \in B_\delta(h_1,g_1)} \hat{\varphi}_P(\xi, h_2, g_2) \leq \tilde{\varphi}_P(\xi, h, g) + \varepsilon/2$, and hence $\tilde{\varphi}_P(\xi, h_1, g_1) \leq \tilde{\varphi}_P(\xi, h, g) + \varepsilon$. This shows that for each $\varepsilon > 0$, there is a $\delta_\varepsilon > 0$ such that for all $(h_1, g_1) \in B_{\delta_\varepsilon/2}(h, g)$, $\tilde{\varphi}_P(\xi, h_1, g_1) \leq \tilde{\varphi}_P(\xi, h, g) + \varepsilon$. **Second**, we prove (ii), that is,

$$\sup_{(h,g) \in \bar{\mathcal{H}} \times \mathcal{G}} \hat{\varphi}_P(\xi, h, g) = \sup_{(h,g) \in \bar{\mathcal{H}} \times \mathcal{G}} \tilde{\varphi}_P(\xi, h, g). \tag{C.30}$$

By the definition of $\tilde{\varphi}_P$, we have $\hat{\varphi}_P(\xi, h, g) \leq \tilde{\varphi}_P(\xi, h, g)$ for all $(h, g) \in \bar{\mathcal{H}} \times \mathcal{G}$, and hence $\sup_{(h,g) \in \bar{\mathcal{H}} \times \mathcal{G}} \hat{\varphi}_P(\xi, h, g) \leq \sup_{(h,g) \in \bar{\mathcal{H}} \times \mathcal{G}} \tilde{\varphi}_P(\xi, h, g)$. Also, by the definition of $\tilde{\varphi}_P$, $\tilde{\varphi}_P(\xi, h, g) \leq \sup_{(h',g') \in \bar{\mathcal{H}} \times \mathcal{G}} \hat{\varphi}_P(\xi, h', g')$ for all (h, g). Thus $\sup_{(h,g) \in \bar{\mathcal{H}} \times \mathcal{G}} \tilde{\varphi}_P(\xi, h, g) \leq \sup_{(h,g) \in \bar{\mathcal{H}} \times \mathcal{G}} \hat{\varphi}_P(\xi, h, g)$, and (C.30) holds. **Similarly**, by the definition of $\tilde{\varphi}_P$, we have that $\hat{\varphi}_P(\xi, h, g) + r_n^{-1}\psi(\xi, h, g) \leq \tilde{\varphi}_P(\xi, h, g) + r_n^{-1}\psi(\xi, h, g)$ for all $(h, g) \in \bar{\mathcal{H}} \times \mathcal{G}$, and hence

$$\sup_{(h,g)\in\bar{\mathcal{H}}\times\mathcal{G}}\{\hat{\varphi}_P(\xi,h,g)+r_n^{-1}\psi(\xi,h,g)\}\leq\sup_{(h,g)\in\bar{\mathcal{H}}\times\mathcal{G}}\{\tilde{\varphi}_P(\xi,h,g)+r_n^{-1}\psi(\xi,h,g)\}.$$

Fix $(h,g)\in\bar{\mathcal{H}}\times\mathcal{G}$. Since $\psi(\xi,\cdot,\cdot)$ is continuous under ρ_P, for every $\varepsilon>0$ there is a $\bar{\delta}>0$ such that

$$\sup_{(h',g')\in B_\delta(h,g)}\{\hat{\varphi}_P(\xi,h',g')+r_n^{-1}\psi(\xi,h,g)-\varepsilon\}\leq\sup_{(h',g')\in B_\delta(h,g)}\{\hat{\varphi}_P(\xi,h',g')+r_n^{-1}\psi(\xi,h',g')\}$$

for all $\delta\leq\bar{\delta}$. By the definition of $\tilde{\varphi}_P$, this implies that

$$\begin{aligned}\tilde{\varphi}_P(\xi,h,g)+r_n^{-1}\psi(\xi,h,g)-\varepsilon&\leq\lim_{\delta\downarrow 0}\sup_{(h',g')\in B_\delta(h,g)}\{\hat{\varphi}_P(\xi,h',g')+r_n^{-1}\psi(\xi,h',g')\}\\&\leq\sup_{(h,g)\in\bar{\mathcal{H}}\times\mathcal{G}}\{\hat{\varphi}_P(\xi,h,g)+r_n^{-1}\psi(\xi,h,g)\}.\end{aligned}$$

Since ε is arbitrary, we have

$$\tilde{\varphi}_P(\xi,h,g)+r_n^{-1}\psi(\xi,h,g)\leq\sup_{(h,g)\in\bar{\mathcal{H}}\times\mathcal{G}}\{\hat{\varphi}_P(\xi,h,g)+r_n^{-1}\psi(\xi,h,g)\}.$$

This holds for all $(h,g)\in\bar{\mathcal{H}}\times\mathcal{G}$, which implies that

$$\sup_{(h,g)\in\bar{\mathcal{H}}\times\mathcal{G}}\{\hat{\varphi}_P(\xi,h,g)+r_n^{-1}\psi(\xi,h,g)\}\geq\sup_{(h,g)\in\bar{\mathcal{H}}\times\mathcal{G}}\{\tilde{\varphi}_P(\xi,h,g)+r_n^{-1}\psi(\xi,h,g)\}.$$

Thus (iii) is proved.

Last, we prove (iv). Since $\varphi_P(\xi,\cdot,\cdot)$ is continuous, we have that

$$\begin{aligned}&\sup_{(\xi,h,g)\in\Xi\times\bar{\mathcal{H}}\times\mathcal{G}}\left|\tilde{\varphi}_P(\xi,h,g)+r_n^{-1}\varphi(\xi,h,g)-\varphi_P(\xi,h,g)\right|\\\leq&\sup_{(\xi,h,g)\in\Xi\times\bar{\mathcal{H}}\times\mathcal{G}}|\hat{\varphi}_P(\xi,h,g)-\varphi_P(\xi,h,g)|+r_n^{-1}\|\varphi\|_\infty.\end{aligned}$$

(iv) follows from the facts that $\hat{\varphi}_P\to\varphi_P$ almost uniformly, as mentioned at the beginning of the proof, and $\|\varphi\|_\infty<\infty$.

Fix $\varepsilon>0$. By property (iv), let $\Omega_0\subset\Omega$ be a measurable set such that $\mathbb{P}(\Omega_0)\geq 1-\varepsilon$ and $\tilde{\varphi}_P+r_n^{-1}\varphi\to\varphi_P$ uniformly on Ω_0 for all $\varphi\in\ell^\infty\left(\Xi\times\bar{\mathcal{H}}\times\mathcal{G}\right)$. Let $\psi_n\in\mathbb{D}_n(\omega_n)$, $\omega_n\in\Omega_0$, and $\psi\in C\left(\Xi\times\bar{\mathcal{H}}\times\mathcal{G}\right)$ be arbitrary maps such that $\psi_n\to\psi$. By property (i) that we proved above, we have that $\Psi\left(\xi,\tilde{\varphi}_P+r_n^{-1}\psi\right)\neq\varnothing$ for all $\omega\in\Omega_0$, all n, and all $\xi\in\Xi$. It is easy to show that because $\psi_n \to \psi$ in $\ell^\infty\left(\Xi \times \bar{\mathcal{H}} \times \mathcal{G}\right)$,

$$\sup_{\xi\in\Xi}\left|\begin{array}{c} \sup_{(h,g)\in\bar{\mathcal{H}}\times\mathcal{G}}\left\{\hat{\varphi}_P(\omega_n)\left(\xi,h,g\right)+r_n^{-1}\psi_n\left(\xi,h,g\right)\right\} \\ -\sup_{(h,g)\in\bar{\mathcal{H}}\times\mathcal{G}}\left\{\hat{\varphi}_P(\omega_n)\left(\xi,h,g\right)+r_n^{-1}\psi\left(\xi,h,g\right)\right\}\end{array}\right|$$
$$\leq r_n^{-1}\sup_{(\xi,h,g)\in\Xi\times\bar{\mathcal{H}}\times\mathcal{G}}\left|\psi_n(\xi,h,g)-\psi(\xi,h,g)\right| = o\left(r_n^{-1}\right).$$

Since $\tilde{\varphi}_P + r_n^{-1}\psi$ converges to φ_P uniformly on Ω_0, by Lemma C.14 there is a sequence $\delta_n \downarrow 0$ such that $\Psi\left(\xi, \tilde{\varphi}_P(\omega) + r_n^{-1}\psi\right) \subset \Psi\left(\xi, \varphi_P\right)^{\delta_n}$ for all $\xi \in \Xi$ and all $\omega \in \Omega_0$. (By Lemma C.14, δ_n does not depend on $\xi \in \Xi$ or on $\omega \in \Omega_0$.) Since $\mathcal{S}(\varphi_P) = 0$ by Lemma C.13, we have that for all $\xi \in \Xi$,

$$\Psi\left(\xi,\varphi_P\right) = \{(h,g)\in\bar{\mathcal{H}}\times\mathcal{G} : \phi_P\left(h,g\right)=0\}. \tag{C.31}$$

By Lemma C.13 and the constructions of $\hat{\varphi}_P$ and $\tilde{\varphi}_P$, we also have that for all ω, $\hat{\varphi}_P \leq 0$ and $\tilde{\varphi}_P \leq 0$ on $\Xi \times \bar{\mathcal{H}} \times \mathcal{G}$, and $\hat{\varphi}_P\left(\xi,\cdot,\cdot\right) = 0$ on $\Psi\left(\xi,\varphi_P\right)$. Thus for every $\xi \in \Xi$,

$$\sup_{(h,g)\in\bar{\mathcal{H}}\times\mathcal{G}}\left\{\hat{\varphi}_P(\omega_n)\left(\xi,h,g\right)+r_n^{-1}\psi\left(\xi,h,g\right)\right\}$$
$$\geq \sup_{(h,g)\in\Psi(\xi,\varphi_P)}\left\{\hat{\varphi}_P(\omega_n)\left(\xi,h,g\right)+r_n^{-1}\psi\left(\xi,h,g\right)\right\} = \sup_{(h,g)\in\Psi(\xi,\varphi_P)} r_n^{-1}\psi\left(\xi,h,g\right).$$

By property (iii) of $\tilde{\varphi}_P$, together with the results shown above, we have that

$$\sup_{\xi\in\Xi}\left|\sup_{(h,g)\in\bar{\mathcal{H}}\times\mathcal{G}}\left\{\hat{\varphi}_P(\omega_n)\left(\xi,h,g\right)+r_n^{-1}\psi\left(\xi,h,g\right)\right\} - \sup_{(h,g)\in\Psi(\xi,\varphi_P)} r_n^{-1}\psi\left(\xi,h,g\right)\right|$$
$$= \sup_{\xi\in\Xi}\left\{\begin{array}{c}\sup_{(h,g)\in\Psi\left(\xi,\tilde{\varphi}_P(\omega_n)+r_n^{-1}\psi\right)}\left\{\tilde{\varphi}_P(\omega_n)\left(\xi,h,g\right)+r_n^{-1}\psi\left(\xi,h,g\right)\right\} \\ -\sup_{(h,g)\in\Psi(\xi,\varphi_P)} r_n^{-1}\psi\left(\xi,h,g\right)\end{array}\right\}$$
$$\leq \sup_{\xi\in\Xi}\left\{\sup_{(h,g)\in\Psi(\xi,\varphi_P)^{\delta_n}}\left\{\tilde{\varphi}_P(\omega_n)\left(\xi,h,g\right)+r_n^{-1}\psi\left(\xi,h,g\right)\right\} - \sup_{(h,g)\in\Psi(\xi,\varphi_P)} r_n^{-1}\psi\left(\xi,h,g\right)\right\}.$$

Then by the definition of $\Psi(\xi,\varphi_P)^{\delta_n}$,

$$\sup_{\xi\in\Xi}\left\{\sup_{(h,g)\in\Psi(\xi,\varphi_P)^{\delta_n}}\left\{\tilde{\varphi}_P(\omega_n)\left(\xi,h,g\right)+r_n^{-1}\psi\left(\xi,h,g\right)\right\} - \sup_{(h,g)\in\Psi(\xi,\varphi_P)} r_n^{-1}\psi\left(\xi,h,g\right)\right\}$$
$$\leq \sup_{\xi\in\Xi}\left\{\sup_{\rho_P((h_1,g_1),(h_2,g_2))\leq\delta_n} r_n^{-1}\left|\psi\left(\xi,h_1,g_1\right)-\psi\left(\xi,h_2,g_2\right)\right|\right\} = o(r_n^{-1}).$$

Finally, combining all the results above, we can conclude that

$$\sup_{\xi\in\Xi}\left|\mathcal{S}\left(\hat{\varphi}_P(\omega_n)+r_n^{-1}\psi_n\right)(\xi)-r_n^{-1}\sup_{(h,g)\in\Psi(\xi,\varphi_P)}\psi\left(\xi,h,g\right)\right|=o\left(r_n^{-1}\right).$$

This implies that

$$\begin{aligned}&\left|g_n(\omega_n)\left(\psi_n\right)-\int_{\Xi}\sup_{(h,g)\in\Psi(\xi,\varphi_P)}\psi\left(\xi,h,g\right)\,\mathrm{d}\nu\left(\xi\right)\right|\\ \leq&\int_{\Xi}\left|r_n\mathcal{S}\left(\hat{\varphi}_P(\omega_n)+r_n^{-1}\psi_n\right)(\xi)-\sup_{(h,g)\in\Psi(\xi,\varphi_P)}\psi\left(\xi,h,g\right)\right|\,\mathrm{d}\nu\left(\xi\right)=o\left(1\right).\end{aligned}$$

■

Proof of Theorem 3.1. By (C.13), $\sqrt{n}(\hat{\phi}_{P_n}-\phi_P)\rightsquigarrow\mathcal{L}'_P(\mathbb{G}_P+Q_0)$, where $\mathcal{L}'_P(\mathbb{G}_P+Q_0)$ is tight as shown in the proof of Lemma 3.1. By Lemma C.11, $\mathcal{M}(\hat{\sigma}_{P_n})\to\mathcal{M}(\sigma_P)$ almost uniformly, and hence this convergence is also in outer probability by Lemma 1.9.3(ii) of van der Vaart and Wellner (1996). By Lemma 1.10.2(iii) of van der Vaart and Wellner (1996), $\mathcal{M}(\hat{\sigma}_{P_n})\rightsquigarrow\mathcal{M}(\sigma_P)$. By Example 1.4.7 (Slutsky's lemma) of van der Vaart and Wellner (1996), we have that $(\sqrt{n}(\hat{\phi}_{P_n}-\phi_P),\mathcal{M}(\hat{\sigma}_{P_n}))\rightsquigarrow(\mathcal{L}'_P(\mathbb{G}_P+Q_0),\mathcal{M}(\sigma_P))$. Let $\ell^\infty(\Xi\times\bar{\mathcal{H}}\times\mathcal{G})^+=\{\psi\in\ell^\infty(\Xi\times\bar{\mathcal{H}}\times\mathcal{G}):\|1/\psi\|_\infty<\infty\}$. Define a map $f:\ell^\infty(\bar{\mathcal{H}}\times\mathcal{G})\times\ell^\infty(\Xi\times\bar{\mathcal{H}}\times\mathcal{G})^+\to\ell^\infty(\Xi\times\bar{\mathcal{H}}\times\mathcal{G})$ by $f(\varphi,\psi)=\varphi/\psi$ for all $(\varphi,\psi)\in\ell^\infty(\bar{\mathcal{H}}\times\mathcal{G})\times\ell^\infty(\Xi\times\bar{\mathcal{H}}\times\mathcal{G})^+$. Clearly, $(\mathcal{L}'_P(\mathbb{G}_P+Q_0),\mathcal{M}(\sigma_P))$ takes its values in $\ell^\infty(\bar{\mathcal{H}}\times\mathcal{G})\times\ell^\infty(\Xi\times\bar{\mathcal{H}}\times\mathcal{G})^+$. It is easy to show that f is continuous under the metric $\|(\varphi,\psi)-(\varphi',\psi')\|=\|\varphi-\varphi'\|_\infty+\|\psi-\psi'\|_\infty$. By Theorem 1.3.6 (continuous mapping) of van der Vaart and Wellner (1996),

$$f(\sqrt{n}(\hat{\phi}_{P_n}-\phi_P),\mathcal{M}(\hat{\sigma}_{P_n}))=\frac{\sqrt{n}(\hat{\phi}_{P_n}-\phi_P)}{\mathcal{M}(\hat{\sigma}_{P_n})}\rightsquigarrow\frac{\mathcal{L}'_P(\mathbb{G}_P+Q_0)}{\mathcal{M}(\sigma_P)}.$$

By Lemma C.13, we have that $\mathcal{I}\circ\mathcal{S}\left(\phi_P/\mathcal{M}\left(\hat{\sigma}_{P_n}\right)\right)=0$. Then by Theorem A.2(ii) and Lemma C.15, together with the continuity of $\mathcal{I}\circ\mathcal{S}_{\Psi(\xi,\varphi_P)}$ under $\|\cdot\|_\infty$, we have

$$\sqrt{n}\left\{\mathcal{I}\circ\mathcal{S}\left(\frac{\hat{\phi}_{P_n}}{\mathcal{M}\left(\hat{\sigma}_{P_n}\right)}\right)-\mathcal{I}\circ\mathcal{S}\left(\frac{\phi_P}{\mathcal{M}\left(\hat{\sigma}_{P_n}\right)}\right)\right\}\rightsquigarrow\mathcal{I}\circ\mathcal{S}_{\Psi(\xi,\varphi_P)}\left(\frac{\mathcal{L}'_P(\mathbb{G}_P+Q_0)}{\mathcal{M}\left(\sigma_P\right)}\right).\tag{C.32}$$

By Lemma C.11, $T_n/n\to\Lambda(P)$ almost uniformly. Then by Lemmas 1.9.3(ii) and 1.10.2(iii), Example 1.4.7 (Slutsky's lemma), and Theorem 1.3.6 (continuous mapping) of van der Vaart and Wellner (1996), together with (C.32), we have that

$$\sqrt{\frac{T_n}{n}}\cdot\sqrt{n}\left\{\mathcal{I}\circ\mathcal{S}\left(\frac{\hat{\phi}_{P_n}}{\mathcal{M}\left(\hat{\sigma}_{P_n}\right)}\right)\right\}\rightsquigarrow\mathcal{I}\circ\mathcal{S}_{\Psi(\xi,\varphi_P)}\left(\frac{\mathbb{G}}{\mathcal{M}(\sigma_P)}\right),$$

where $\mathbb{G} = \sqrt{\Lambda(P)}\mathcal{L}'_P(\mathbb{G}_P + Q_0)$ as in Lemma 3.1. By Lemma C.13, we have that $\Psi(\xi, \varphi_P) = \Psi_{\bar{\mathcal{H}}\times\mathcal{G}}$ defined by (25) for all $\xi \in \Xi$ under the assumptions.

If $\mathcal{D}$ is a finite set with $\mathcal{D} = \{d_1, \ldots, d_J\}$, then under null $\mathcal{I} \circ \mathcal{S}_{\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}}(\mathbb{G}/\mathcal{M}(\sigma_P)) = \mathcal{I} \circ \mathcal{S}_{\Psi_{\mathcal{H}\times\mathcal{G}}}(\mathbb{G}/\mathcal{M}(\sigma_P))$ almost surely, because it can be shown that in this special case $\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}$ is equal to the closure of $\Psi_{\mathcal{H}\times\mathcal{G}}$ in $\bar{\mathcal{H}} \times \mathcal{G}$ under ρ_P and $\mathbb{G}/\mathcal{M}(\sigma_P)$ is continuous under ρ_P almost surely for every fixed ξ. We summarize this in the following.

By Lemma C.13, $\phi_P(h, g) \leq 0$ for all $(h, g) \in \bar{\mathcal{H}}\times\mathcal{G}$, and there exists $(h^0, g^0) \in \bar{\mathcal{H}}\times\mathcal{G}$ with $g^0 = (g_1^0, g_2^0)$ such that $\phi_P(h^0, g^0) = 0$. **First**, we show that if $h^0 = (-1)^d \cdot 1_{A\times\{d\}\times\mathbb{R}}$, where $d \in \{0, 1\}$ and A is a half-closed interval or an open interval, then for every closed interval B such that $B \subset A$, we have that $\phi_P(\tilde{h}, g^0) = 0$ with $\tilde{h} = (-1)^d \cdot 1_{B\times\{d\}\times\mathbb{R}}$. Suppose, by way of contradiction, that $A = (a_1, a_2)$ and $B = [b_1, b_2]$ with $a_1 < b_1$, $a_2 > b_2$, and $\phi_P(\tilde{h}, g^0) < 0$ with $\tilde{h} = (-1)^d \cdot 1_{B\times\{d\}\times\mathbb{R}}$. Let $h_L = (-1)^d \cdot 1_{(a_1,b_1)\times\{d\}\times\mathbb{R}}$ and $h_R = (-1)^d \cdot 1_{(b_2,a_2)\times\{d\}\times\mathbb{R}}$. Then by the definition of ϕ_P,

$$\begin{aligned}\phi_P(h^0, g^0) &= \frac{P(h^0 \cdot g_2^0)}{P(g_2^0)} - \frac{P(h^0 \cdot g_1^0)}{P(g_1^0)} = \frac{P((h_L + \tilde{h} + h_R) \cdot g_2^0)}{P(g_2^0)} - \frac{P((h_L + \tilde{h} + h_R) \cdot g_1^0)}{P(g_1^0)} \\ &= \phi_P(\tilde{h}, g^0) + \phi_P(h_L, g^0) + \phi_P(h_R, g^0).\end{aligned}$$

Since $\phi_P(h^0, g^0) = 0$ but $\phi_P(\tilde{h}, g^0) < 0$, we have $\phi_P(h_L, g^0) + \phi_P(h_R, g^0) > 0$. This implies that either $\phi_P(h_L, g^0) > 0$ or $\phi_P(h_R, g^0) > 0$. However, since $(h_L, g^0), (h_R, g^0) \in \bar{\mathcal{H}} \times \mathcal{G}$, Lemma C.13 shows that both $\phi_P(h_L, g^0)$ and $\phi_P(h_R, g^0)$ are nonpositive. This is a contradiction. When A is a half-closed interval, we can show analogously that the claim is true. **Second**, we show that if $h^0 = 1_{\mathbb{R}\times C\times\mathbb{R}}$ with $C = (-\infty, c)$ for some $c \in \mathbb{R}$, then there is a sequence of sets $C_k = (-\infty, c_k]$ with $c_k \uparrow c$ such that $\phi_P(h^k, g^0) = 0$ with $h^k = 1_{\mathbb{R}\times C_k\times\mathbb{R}}$. By assumption, $\mathcal{D}$ is a finite set. Under Assumption 3.1, D is a discrete random variable with $D \in \mathcal{D}$ under P_n. Then $D \in \mathcal{D}$ under P by Lemma C.11, and the claim holds.

The above results imply that $\Psi_{\bar{\mathcal{H}}\times\mathcal{G}} \subset \overline{\Psi_{\mathcal{H}\times\mathcal{G}}}$, where $\overline{\Psi_{\mathcal{H}\times\mathcal{G}}}$ is the closure of $\Psi_{\mathcal{H}\times\mathcal{G}}$ in $\bar{\mathcal{H}} \times \mathcal{G}$ under ρ_P. By (25) and Lemma C.12, $\Psi_{\bar{\mathcal{H}}\times\mathcal{G}} = \overline{\Psi_{\mathcal{H}\times\mathcal{G}}}$. By Lemma 3.1, $\mathbb{G}$ almost surely has a continuous path under ρ_P. By Lemma C.12, σ_P is continuous under ρ_P. The result follows from the continuity of $\mathbb{G}/\mathcal{M}(\sigma_P)$ under ρ_P for every fixed $\xi \in \Xi$. ■

Remark C.3 *If the H_0 in (15) is true with $Q = P_n$ for all n, we have that $\mathcal{S}(\phi_P/\mathcal{M}(\sigma_P)) = 0$ (see Lemma C.13). Thus it suffices to find the asymptotic distribution of*

$$\sqrt{n}\mathcal{I} \circ \mathcal{S}_{\mathcal{H}\times\mathcal{G}}\left(\frac{\hat{\phi}_{P_n}}{\mathcal{M}(\hat{\sigma}_{P_n})}\right) = \sqrt{n}\left\{\mathcal{I} \circ \mathcal{S}\left(\frac{\hat{\phi}_{P_n}}{\mathcal{M}(\hat{\sigma}_{P_n})}\right) - \mathcal{I} \circ \mathcal{S}\left(\frac{\phi_P}{\mathcal{M}(\sigma_P)}\right)\right\}. \qquad \text{(C.33)}$$

If we can find the asymptotic distribution of $\sqrt{n}(\hat{\phi}_{P_n}/\mathcal{M}(\hat{\sigma}_{P_n}) - \phi_P/\mathcal{M}(\sigma_P))$ and the "derivative" of $\mathcal{I} \circ \mathcal{S}$ (see, for example, the definition of Hadamard directional derivative in Shapiro

(1990) and Fang and Santos (2018)), then by the delta method of Fang and Santos (2018), it is straightforward to obtain the asymptotic distribution of (C.33). However, establishing the limiting distribution of $\sqrt{n}(\hat{\phi}_{P_n}/\mathcal{M}(\hat{\sigma}_{P_n}) - \phi_P/\mathcal{M}(\sigma_P))$ is technically tricky. By the constructions of ϕ_P and σ_P, we can view $\phi_P/\mathcal{M}(\sigma_P)$ as a map of P. Specifically, let $\mathcal{V}_0 = \{v : v = h \cdot g_l \text{ or } v = h^2 \cdot g_l \text{ for some } h \in \bar{\mathcal{H}} \text{ and } g_l \in \mathcal{G}_K\}$ and $\mathbb{D}_Q = \{Q \in \ell^\infty(\mathcal{V}_0 \cup \mathcal{G}_K) : Q(h \cdot g_l)/Q(g_l) \text{ and } Q(h^2 \cdot g_l)/Q(g_l) \text{ exist for all } h \in \bar{\mathcal{H}} \text{ and } g_l \in \mathcal{G}_K\}$. Then we extend the definitions of ϕ_Q and σ_Q for all $Q \in \mathcal{P}$, that is, the ϕ_Q defined in (14) and the σ_Q defined in (19), to all $Q \in \mathbb{D}_Q$. Clearly, $\mathcal{P} \subset \mathbb{D}_Q$ by (13). Define a map $\mathcal{T} : \mathbb{D}_Q \to \ell^\infty(\Xi \times \bar{\mathcal{H}} \times \mathcal{G})$ by

$$\mathcal{T}(Q)(\xi, h, g) = \frac{\phi_Q(h, g)}{\mathcal{M}(\sigma_Q)(\xi, h, g)}$$

for all $Q \in \mathbb{D}_Q$ and $(\xi, h, g) \in \Xi \times \bar{\mathcal{H}} \times \mathcal{G}$. Now we have that $\mathcal{T}(P) = \phi_P/\mathcal{M}(\sigma_P)$ and $\mathcal{T}(\hat{P}_n) = \hat{\phi}_{P_n}/\mathcal{M}(\hat{\sigma}_{P_n})$. Suppose we have weak convergence of $\sqrt{n}(\hat{P}_n - P)$ in some suitable space. Then if $\mathcal{T}$ is Hadamard (directionally) differentiable, by delta method we can establish weak convergence of

$$\sqrt{n}\left(\frac{\hat{\phi}_{P_n}}{\mathcal{M}(\hat{\sigma}_{P_n})} - \frac{\phi_P}{\mathcal{M}(\sigma_P)}\right) = \sqrt{n}\left(\mathcal{T}(\hat{P}_n) - \mathcal{T}(P)\right). \tag{C.34}$$

Unfortunately, however, $\mathcal{T}$ is nondifferentiable, because of the nondifferentiability of the $\mathcal{M}$ defined in (C.1) (to the best of our knowledge, the directional derivative of $\mathcal{M}$ may not exist even when Ξ is a singleton), and hence it is not straightforward to show the convergence of $\sqrt{n}(\mathcal{T}(\hat{P}_n) - \mathcal{T}(P))$. The random denominator problem also arises in other testing issues. See, for example, Bugni et al. (2017). If the random denominator does not need to be bounded away from 0 in the test statistic as in Bugni et al. (2017), we will not have the undifferentiability issue caused by $\mathcal{M}$. Inspired by Kitagawa (2015), with the asymptotic distribution of $\sqrt{n}(\hat{\phi}_{P_n}/\mathcal{M}(\hat{\sigma}_{P_n}) - \phi_P/\mathcal{M}(\hat{\sigma}_{P_n}))$ (which can be obtained by using Slutsky's theorem), we can instead establish the asymptotic distribution of

$$\sqrt{n}\left\{\mathcal{I} \circ \mathcal{S}\left(\frac{\hat{\phi}_{P_n}}{\mathcal{M}(\hat{\sigma}_{P_n})}\right) - \mathcal{I} \circ \mathcal{S}\left(\frac{\phi_P}{\mathcal{M}(\hat{\sigma}_{P_n})}\right)\right\}, \tag{C.35}$$

where $\mathcal{S}(\phi_P/\mathcal{M}(\hat{\sigma}_{P_n})) = 0$ by Lemma C.13 if the H_0 in (15) is true with $Q = P_n$ for all n. However, existing delta methods cannot be used to establish the asymptotic distribution of (C.35) either. Since $\phi_P/\mathcal{M}(\hat{\sigma}_{P_n})$ is a random element, delta methods such as Theorem 3.9.4 or Theorem 3.9.5 of van der Vaart and Wellner (1996), or Theorem 2.1 of Fang and Santos (2018), do not work in this case. To overcome the technical complications due to the random element $\phi_P/\mathcal{M}(\hat{\sigma}_{P_n})$, we provide the extended continuous mapping theorem and the extended *delta method elaborated by Theorems A.1 and A.2, respectively.*

We now introduce the notation for the bootstrap elements. Let $(W_{n1}, \ldots, W_{nn})$ be a vector of random multinomial weights independent of $\{(Y_i, D_i, Z_i)\}_{i=1}^n$ for all n. As defined in (16), $\hat{P}_n$ is the empirical measure of an i.i.d. sample $\{(Y_i, D_i, Z_i)\}_{i=1}^n$ from probability distribution P_n. Given the sample values, the $\{(\hat{Y}_i, \hat{D}_i, \hat{Z}_i)\}_{i=1}^n$ introduced in Section 3.1.1 is an i.i.d. sample from $\hat{P}_n$. We can write the empirical measure of $\{(\hat{Y}_i, \hat{D}_i, \hat{Z}_i)\}_{i=1}^n$, given sample $\{(Y_i, D_i, Z_i)\}_{i=1}^n$, as $\hat{P}_n^B = n^{-1}\sum_{i=1}^n W_{ni}\delta_{(Y_i,D_i,Z_i)}$, where $\delta_{(Y_i,D_i,Z_i)}$ is a Dirac measure centered at (Y_i, D_i, Z_i). Given the $\hat{\phi}_{P_n}^B$, T_n^B, and $\hat{\sigma}_{P_n}^B$ defined in Section 3.1.1, $\hat{\phi}_{P_n}^B/\mathcal{M}(\hat{\sigma}_{P_n}^B)$ is a map of $\{(Y_i, D_i, Z_i, W_{ni})\}_{i=1}^n$ to the space $\ell^\infty\left(\Xi \times \bar{\mathcal{H}} \times \mathcal{G}\right)$.

We follow Section 3.6 of van der Vaart and Wellner (1996) and (A.1) to define the conditional outer expectations. When we compute the outer expectations as in (A.1), independence is understood in terms of a product space. Under Assumptions 3.1 and 3.2, each term (Y_i, D_i, Z_i) of the sequence $\{(Y_i, D_i, Z_i)\}_{i=1}^\infty$ has probability distribution P. Let $\{(Y_i, D_i, Z_i)\}_{i=1}^\infty$ be the coordinate projections on the first ∞ coordinates of the product space $((\mathbb{R}^3)^\infty, \mathcal{B}_{\mathbb{R}^3}^\infty, P^\infty) \times (\mathcal{W}, \mathcal{C}, P_W)$, and let the multinomial vectors W depend on the last factor only. For each real-valued map T on $((\mathbb{R}^3)^\infty, \mathcal{B}_{\mathbb{R}^3}^\infty, P^\infty) \times (\mathcal{W}, \mathcal{C}, P_W)$, we can take $(\Omega_1, \mathcal{A}_1, \mathbb{P}_1) = ((\mathbb{R}^3)^\infty, \mathcal{B}_{\mathbb{R}^3}^\infty, P^\infty)$ and $(\Omega_2, \mathcal{A}_2, \mathbb{P}_2) = (\mathcal{W}, \mathcal{C}, P_W)$ and define a real-valued map $E_W^*[T]$ on $((\mathbb{R}^3)^\infty, \mathcal{B}_{\mathbb{R}^3}^\infty, P^\infty)$ by

$$E_W^*[T](\{(Y_i, D_i, Z_i)\}_{i=1}^\infty) = E_2^*[T](\{(Y_i, D_i, Z_i)\}_{i=1}^\infty) \tag{C.36}$$

for each sequence $\{(Y_i, D_i, Z_i)\}_{i=1}^\infty \in (\mathbb{R}^3)^\infty$, where $E_2^*[T]$ is defined as in (A.1). We call the left-hand side of (C.36) the conditional outer expectation of T given the sequence $\{(Y_i, D_i, Z_i)\}_{i=1}^\infty$. Since $E_W^*[T]$ is a real-valued map on $((\mathbb{R}^3)^\infty, \mathcal{B}_{\mathbb{R}^3}^\infty, P^\infty)$, we can compute its outer and inner integrals (expectations) with respect to $((\mathbb{R}^3)^\infty, \mathcal{B}_{\mathbb{R}^3}^\infty, P^\infty)$. For simplicity of notation, we write them as $E^*[E_W^*[T]]$ and $E_*[E_W^*[T]]$, respectively.

If $T(\{(Y_i, D_i, Z_i)\}_{i=1}^\infty, \cdot)$ is a measurable integrable map on $(\mathcal{W}, \mathcal{C}, P_W)$ for every given sequence $\{(Y_i, D_i, Z_i)\}_{i=1}^\infty$, we write $E_W[T]$ for $E_W^*[T]$ and call $E_W[T](\{(Y_i, D_i, Z_i)\}_{i=1}^\infty)$ the conditional expectation of T given the sequence $\{(Y_i, D_i, Z_i)\}_{i=1}^\infty$. The conditional inner expectation is defined analogously. If $\mathbb{D}$ is a metric space with metric d, we define

$$\mathrm{BL}_1(\mathbb{D}) = \{f : \mathbb{D} \to \mathbb{R} : \|f\|_\infty \leq 1, |f(x_1) - f(x_2)| \leq d(x_1, x_2) \text{ for all } x_1, x_2 \in \mathbb{D}\}.$$

Lemma C.16 *Suppose Assumptions 3.1 and 3.2 hold.*

(i) $\sqrt{T_n^B}(\hat{\phi}^B_{P_n} - \hat{\phi}_{P_n})/\mathcal{M}(\hat{\sigma}^B_{P_n})$ *satisfies*

$$\sup_{f \in \mathrm{BL}_1(\ell^\infty(\Xi \times \bar{\mathcal{H}} \times \mathcal{G}))} \left| E_W \left[f\left(\frac{\sqrt{T_n^B}(\hat{\phi}^B_{P_n} - \hat{\phi}_{P_n})}{\mathcal{M}(\hat{\sigma}^B_{P_n})} \right) \right] - E\left[f\left(\frac{\mathbb{G}_0}{\mathcal{M}(\sigma_P)} \right) \right] \right| \to 0 \quad \text{(C.37)}$$

in outer probability, where $\mathbb{G}_0 = \sqrt{\Lambda(P)} \cdot \mathcal{L}'_P(\mathbb{G}_P)$ *is tight and* $\mathbb{G}_P$ *is as in Lemma C.10;*

(ii) $\sqrt{T_n^B}(\hat{\phi}^B_{P_n} - \hat{\phi}_{P_n})/\mathcal{M}(\hat{\sigma}^B_{P_n}) \rightsquigarrow \mathbb{G}_0/\mathcal{M}(\sigma_P)$;[9]

(iii) For each continuous, bounded $f : \ell^\infty(\Xi \times \bar{\mathcal{H}} \times \mathcal{G}) \to \mathbb{R}$, $f(\sqrt{T_n^B}(\hat{\phi}^B_{P_n} - \hat{\phi}_{P_n})/\mathcal{M}(\hat{\sigma}^B_{P_n}))$ *is a measurable function of* $\{W_{ni}\}_{i=1}^n$ *for every given sequence* $\{(Y_i, D_i, Z_i)\}_{i=1}^\infty$.

Proof of Lemma C.16. (i). To explore the conditional property of the bootstrap element $\sqrt{T_n^B}(\hat{\phi}^B_{P_n} - \hat{\phi}_{P_n})/\mathcal{M}(\hat{\sigma}^B_{P_n})$, we consider the entire sequence $\{(Y_i, D_i, Z_i)\}_{i=1}^\infty$.[10] Each term (Y_i, D_i, Z_i) in $\{(Y_i, D_i, Z_i)\}_{i=1}^\infty$ has probability distribution P under Assumptions 3.1 and 3.2. Now the $\hat{P}_n$ defined in (16) can be viewed as being computed with the first n elements of $\{(Y_i, D_i, Z_i)\}_{i=1}^\infty$ that are distributed according to P. By Lemma C.5, $\sqrt{n}(\hat{P}_n - P) \rightsquigarrow \mathbb{G}_P$ under P, where $\mathbb{G}_P$ is the limit shown in Lemma C.10. By the construction of $\tilde{\mathcal{V}}$ in (C.6), $F = 1$ is an envelope function of $\tilde{\mathcal{V}}$ and $P^*(\sup_{v \in \tilde{\mathcal{V}}} |v - P(v)|^2) < \infty$, where P^* is the outer probability measure of P. By Lemma C.5, $\tilde{\mathcal{V}}$ is Donsker. By Theorem 3.6.2 of van der Vaart and Wellner (1996), we have that

$$\sup_{f \in \mathrm{BL}_1(\ell^\infty(\tilde{\mathcal{V}}))} |E_W[f\{\sqrt{n}(\hat{P}^B_n - \hat{P}_n)\}] - E[f(\mathbb{G}_P)]| \to 0 \quad \text{(C.38)}$$

outer almost surely[11] and

$$E_W[f\{\sqrt{n}(\hat{P}^B_n - \hat{P}_n)\}^*] - E_W[f\{\sqrt{n}(\hat{P}^B_n - \hat{P}_n)\}_*] \to 0 \quad \text{(C.39)}$$

almost surely for every $f \in \mathrm{BL}_1(\ell^\infty(\tilde{\mathcal{V}}))$. Here, the asterisks denote the measurable cover functions with respect to $\{(Y_i, D_i, Z_i)\}_{i=1}^\infty$ and W jointly. Then by Lemmas C.9, C.5, and C.6 in this paper, and Theorem 3.9.13 of van der Vaart and Wellner (1996), we have

$$\sup_{f \in \mathrm{BL}_1(\ell^\infty(\bar{\mathcal{H}} \times \mathcal{G}))} |E_W[f\{\sqrt{n}(\mathcal{L}(\hat{P}^B_n) - \mathcal{L}(\hat{P}_n))\}] - E[f(\mathcal{L}'_P(\mathbb{G}_P))]| \to 0 \quad \text{(C.40)}$$

[9] This implies that $\sqrt{T_n^B}(\hat{\phi}^B_{P_n} - \hat{\phi}_{P_n})/\mathcal{M}(\hat{\sigma}^B_{P_n})$ is asymptotically measurable jointly in $\{(Y_i, D_i, Z_i)\}_{i=1}^\infty$ and W by Lemma 1.3.8 of van der Vaart and Wellner (1996).

[10] We follow Section 3.6 of van der Vaart and Wellner (1996) to obtain the conditional property of the bootstrap element $\sqrt{T_n^B}(\hat{\phi}^B_{P_n} - \hat{\phi}_{P_n})/\mathcal{M}(\hat{\sigma}^B_{P_n})$ given the entire sequence $\{(Y_i, D_i, Z_i)\}_{i=1}^\infty$.

[11] As discussed in van der Vaart and Wellner (1996, p. 183), $f\{\sqrt{n}(\hat{P}^B_n - \hat{P}_n)\}$ is measurable as a function of the random weights given the values of the sample. Thus we use the conditional expectation $E_W[f\{\sqrt{n}(\hat{P}^B_n - \hat{P}_n)\}]$ in (C.38). Similarly, we use the conditional expectation $E_W[f\{\sqrt{n}(\mathcal{L}(\hat{P}^B_n) - \mathcal{L}(\hat{P}_n))\}]$ in (C.40).

outer almost surely and

$$E_W[f\{\sqrt{n}(\mathcal{L}(\hat{P}_n^B)-\mathcal{L}(\hat{P}_n))\}^*]-E_W[f\{\sqrt{n}(\mathcal{L}(\hat{P}_n^B)-\mathcal{L}(\hat{P}_n))\}_*]\to 0 \tag{C.41}$$

almost surely for every $f\in \mathrm{BL}_1(\ell^\infty(\bar{\mathcal{H}}\times\mathcal{G}))$. The outer almost sure convergence in (C.40) implies that the weak convergence $\sqrt{n}(\mathcal{L}(\hat{P}_n^B)-\mathcal{L}(\hat{P}_n))\rightsquigarrow\mathcal{L}'_P(\mathbb{G}_P)$ holds for almost every given sequence $\{(Y_i,D_i,Z_i)\}_{i=1}^\infty$. By Lemma C.6 in this paper, and Lemmas 1.9.2 and 1.9.3 of van der Vaart and Wellner (1996), we have that $\|\hat{P}_n^B-\hat{P}_n\|_\infty\to 0$ outer almost surely for almost every given sequence $\{(Y_i,D_i,Z_i)\}_{i=1}^\infty$. By Lemma C.6 again, $\|\hat{P}_n-P\|_\infty\to 0$ for almost every sequence $\{(Y_i,D_i,Z_i)\}_{i=1}^\infty$. Thus now we have that $\|\hat{P}_n^B-P\|_\infty\le\|\hat{P}_n^B-\hat{P}_n\|_\infty+\|\hat{P}_n-P\|_\infty\to 0$ outer almost surely for almost every given sequence $\{(Y_i,D_i,Z_i)\}_{i=1}^\infty$. This implies that $\|\hat{\sigma}_{P_n}^B-\sigma_P\|_\infty\to 0$ and $T_n^B/n\to\Lambda(P)$ outer almost surely for almost every given sequence $\{(Y_i,D_i,Z_i)\}_{i=1}^\infty$. This, together with (C.40), and Lemmas 1.9.2(i) and 1.10.2(iii), Example 1.4.7 (Slutsky's lemma), and Theorem 1.3.6 (continuous mapping) of van der Vaart and Wellner (1996), implies that $\sqrt{T_n^B}(\mathcal{L}(\hat{P}_n^B)-\mathcal{L}(\hat{P}_n))/\mathcal{M}(\hat{\sigma}_{P_n}^B)\rightsquigarrow\mathbb{G}_0/\mathcal{M}(\sigma_P)$ for almost every given sequence $\{(Y_i,D_i,Z_i)\}_{i=1}^\infty$. Since $\mathbb{G}_P$ is tight, $\mathbb{G}_0$ is tight by (C.10).

(ii). By (C.41) and Theorem 2.37 of Folland (1999) (Fubini), together with the dominated convergence theorem and Lemma 1.2.1 of van der Vaart and Wellner (1996),

$$E^*[f\{\sqrt{n}(\mathcal{L}(\hat{P}_n^B)-\mathcal{L}(\hat{P}_n))\}]-E_*[f\{\sqrt{n}(\mathcal{L}(\hat{P}_n^B)-\mathcal{L}(\hat{P}_n))\}]\to 0 \tag{C.42}$$

for every $f\in \mathrm{BL}_1(\ell^\infty(\bar{\mathcal{H}}\times\mathcal{G}))$. By (C.40), together with the definition of outer almost sure convergence (Definition 1.9.1(iii) of van der Vaart and Wellner (1996)), we have that for every function $f\in \mathrm{BL}_1(\ell^\infty(\bar{\mathcal{H}}\times\mathcal{G}))$,

$$|E_W[f\{\sqrt{n}(\mathcal{L}(\hat{P}_n^B)-\mathcal{L}(\hat{P}_n))\}]-E[f(\mathcal{L}'_P(\mathbb{G}_P))]|^*\to 0 \tag{C.43}$$

almost surely. Thus by (C.43), together with Lemma 1.2.2(iii) of van der Vaart and Wellner (1996), we have that

$$|(E_W[f\{\sqrt{n}(\mathcal{L}(\hat{P}_n^B)-\mathcal{L}(\hat{P}_n))\}])^*-E[f(\mathcal{L}'_P(\mathbb{G}_P))]|\to 0 \tag{C.44}$$

almost surely for every $f\in \mathrm{BL}_1(\ell^\infty(\bar{\mathcal{H}}\times\mathcal{G}))$. By Lemma 1.2.6 (Fubini's theorem) of van der Vaart and Wellner (1996),

$$\begin{aligned}E^*[f\{\sqrt{n}(\mathcal{L}(\hat{P}_n^B)-\mathcal{L}(\hat{P}_n))\}]&\ge E^*[E_W[f\{\sqrt{n}(\mathcal{L}(\hat{P}_n^B)-\mathcal{L}(\hat{P}_n))\}]]\\&\ge E_*[f\{\sqrt{n}(\mathcal{L}(\hat{P}_n^B)-\mathcal{L}(\hat{P}_n))\}].\end{aligned} \tag{C.45}$$

Then by Lemma 1.2.1 of van der Vaart and Wellner (1996) and (C.42), we have that

$$E^*[f\{\sqrt{n}(\mathcal{L}(\hat{P}_n^B)-\mathcal{L}(\hat{P}_n))\}]=E[(E_W[f\{\sqrt{n}(\mathcal{L}(\hat{P}_n^B)-\mathcal{L}(\hat{P}_n))\}])^*]+o(1). \tag{C.46}$$

Now with (C.44) we can conclude that

$$\begin{aligned}&|E^*[f\{\sqrt{n}(\mathcal{L}(\hat{P}_n^B)-\mathcal{L}(\hat{P}_n))\}]-E[f(\mathcal{L}'_P(\mathbb{G}_P))]|\\ =&|E[(E_W[f\{\sqrt{n}(\mathcal{L}(\hat{P}_n^B)-\mathcal{L}(\hat{P}_n))\}])^*]+o(1)-E[f(\mathcal{L}'_P(\mathbb{G}_P))]|\\ \le& E[|(E_W[f\{\sqrt{n}(\mathcal{L}(\hat{P}_n^B)-\mathcal{L}(\hat{P}_n))\}])^*-E[f(\mathcal{L}'_P(\mathbb{G}_P))]|]+o(1)\to 0\end{aligned}$$

for every $f\in \mathrm{BL}_1(\ell^\infty(\bar{\mathcal{H}}\times\mathcal{G}))$, where the equality is from (C.46) and the convergence is by the dominated convergence theorem together with the almost sure convergence in (C.44). This implies that $\sqrt{n}(\mathcal{L}(\hat{P}_n^B)-\mathcal{L}(\hat{P}_n))\rightsquigarrow\mathcal{L}'_P(\mathbb{G}_P)$ unconditionally. Similarly, by (C.38) and (C.39) we can easily show that $\sqrt{n}(\hat{P}_n^B-\hat{P}_n)\rightsquigarrow\mathbb{G}_P$ unconditionally. Thus we can conclude that $\hat{P}_n^B-\hat{P}_n\to 0$ in outer probability by Lemma 1.10.2(iii) of van der Vaart and Wellner (1996). By Lemma C.6 in this paper and Lemmas 1.9.3 and 1.2.2(i) of van der Vaart and Wellner (1996), we have that $\hat{P}_n^B\to P$ in outer probability, and hence $T_n^B/n\to\Lambda(P)$ and $\mathcal{M}(\hat{\sigma}_{P_n}^B)\to\mathcal{M}(\sigma_P)$ in outer probability by Theorem 1.9.5 (continuous mapping) of van der Vaart and Wellner (1996). By Lemma 1.10.2(iii), Example 1.4.7 (Slutsky's lemma), and Theorem 1.3.6 (continuous mapping) of van der Vaart and Wellner (1996), $\sqrt{T_n^B}(\mathcal{L}(\hat{P}_n^B)-\mathcal{L}(\hat{P}_n))/\mathcal{M}(\hat{\sigma}_{P_n}^B)\rightsquigarrow\mathbb{G}_0/\mathcal{M}(\sigma_P)$ unconditionally. This verifies (ii) of the Lemma.

(iii). This claim holds naturally under our constructions. ■

To explore the property of the bootstrap test statistic, we introduce the following notation. For all sets $A_1,A_2\subset\bar{\mathcal{H}}\times\mathcal{G}$, define $\overrightarrow{d_H}(A_1,A_2)=\sup_{a\in A_1}\inf_{b\in A_2}\rho_P(a,b)$ and

$$d_H(A_1,A_2)=\max\left\{\overrightarrow{d_H}(A_1,A_2),\overrightarrow{d_H}(A_2,A_1)\right\}.$$

Also, define

$$\widehat{\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}}=\left\{(h,g)\in\bar{\mathcal{H}}\times\mathcal{G}:\sqrt{T_n}\left|\frac{\hat{\phi}_{P_n}(h,g)}{\mathcal{M}(\hat{\sigma}_{P_n})(\xi_0,h,g)}\right|\le\tau_n\right\}, \tag{C.47}$$

where ξ_0 and τ_n are as in (27). Notice the difference between $\widehat{\Psi_{\mathcal{H}\times\mathcal{G}}}$ in (27) and $\widehat{\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}}$ in (C.47). Clearly, $\widehat{\Psi_{\mathcal{H}\times\mathcal{G}}}\subset\widehat{\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}}$.

Lemma C.17 *Under Assumptions 3.1 and 3.2, if the H_0 in (15) is true with $Q=P_n$ for all n, then $d_H(\widehat{\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}},\Psi_{\bar{\mathcal{H}}\times\mathcal{G}})\to 0$ in outer probability, where $\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}$ is defined as in (25).*

Proof of Lemma C.17. First, under the assumptions, we have that for all $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}^* \left(\overrightarrow{d_H} \left(\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}, \widehat{\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}} \right) > \varepsilon \right) \leq \lim_{n\to\infty} \mathbb{P}^* \left(\Psi_{\bar{\mathcal{H}}\times\mathcal{G}} \backslash \widehat{\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}} \neq \varnothing \right)$$
$$\leq \lim_{n\to\infty} \mathbb{P}^* \left(\sup_{(h,g)\in\bar{\mathcal{H}}\times\mathcal{G}} \sqrt{T_n} \left| \frac{\hat{\phi}_{P_n}(h,g) - \phi_P(h,g)}{\xi_0 \vee \hat{\sigma}_{P_n}(h,g)} \right| > \tau_n \right).$$

By Lemma 3.1, $\sqrt{T_n}(\hat{\phi}_{P_n} - \phi_P) \rightsquigarrow \mathbb{G}$. By Lemma C.11, $\hat{\sigma}_{P_n} \to \sigma_P$ almost uniformly, which implies that $\hat{\sigma}_{P_n} \rightsquigarrow \sigma_P$ by Lemmas 1.9.3(ii) and 1.10.2(iii) of van der Vaart and Wellner (1996). Thus by Example 1.4.7 (Slutsky's lemma) and Theorem 1.3.6 (continuous mapping) of van der Vaart and Wellner (1996),

$$\sup_{(h,g)\in\bar{\mathcal{H}}\times\mathcal{G}} \sqrt{T_n} \left| \frac{\hat{\phi}_{P_n}(h,g) - \phi_P(h,g)}{\xi_0 \vee \hat{\sigma}_{P_n}(h,g)} \right| \rightsquigarrow \sup_{(h,g)\in\bar{\mathcal{H}}\times\mathcal{G}} \left| \frac{\mathbb{G}(h,g)}{\xi_0 \vee \sigma_P(h,g)} \right|.$$

Since $\tau_n \to \infty$, we have that $\lim_{n\to\infty} \mathbb{P}^*(\overrightarrow{d_H}(\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}, \widehat{\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}}) > \varepsilon) = 0$.

Next, consider $\overrightarrow{d_H}(\widehat{\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}}, \Psi_{\bar{\mathcal{H}}\times\mathcal{G}})$. Define

$$d((h,g), A) = \inf_{(h',g')\in A} \rho_P((h,g),(h',g'))$$

for all $(h,g) \in \bar{\mathcal{H}} \times \mathcal{G}$ and all subsets $A \subset \bar{\mathcal{H}} \times \mathcal{G}$. For each $\varepsilon > 0$, define

$$\tilde{D}_\varepsilon = \left\{ (h,g) \in \bar{\mathcal{H}} \times \mathcal{G} : d\left((h,g), \Psi_{\bar{\mathcal{H}}\times\mathcal{G}}\right) \geq \varepsilon \right\}.$$

The product space $\bar{\mathcal{H}} \times \mathcal{G}$ is compact under ρ_P by Lemma C.8. Suppose $\{(h_n, g_n)\}_n \subset \tilde{D}_\varepsilon$ such that $(h_n, g_n) \to (h,g)$ for some $(h,g) \in \bar{\mathcal{H}} \times \mathcal{G}$. Then

$$d\left((h,g), \Psi_{\bar{\mathcal{H}}\times\mathcal{G}}\right) = \inf_{(h',g')\in\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}} \rho_P\left((h,g),(h',g')\right)$$
$$\geq \inf_{(h',g')\in\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}} \rho_P\left((h_n,g_n),(h',g')\right) - \rho_P\left((h,g),(h_n,g_n)\right) \geq \varepsilon - \rho_P\left((h,g),(h_n,g_n)\right),$$

which is true for all n. Letting $n \to \infty$ gives $d((h,g), \Psi_{\bar{\mathcal{H}}\times\mathcal{G}}) \geq \varepsilon$. This implies that $\tilde{D}_\varepsilon$ is closed in $\bar{\mathcal{H}} \times \mathcal{G}$, which is compact, and thus $\tilde{D}_\varepsilon$ is compact. If $\tilde{D}_\varepsilon = \varnothing$, then clearly

$$\lim_{n\to\infty} \mathbb{P}^* \left(\overrightarrow{d_H} \left(\widehat{\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}}, \Psi_{\bar{\mathcal{H}}\times\mathcal{G}} \right) > \varepsilon \right)$$
$$= \lim_{n\to\infty} \mathbb{P}^* \left(\sup_{(h,g)\in\widehat{\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}}} \inf_{(h',g')\in\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}} \rho_P\left((h,g),(h',g')\right) > \varepsilon \right) = 0.$$

If $\tilde{D}_\varepsilon \neq \varnothing$, then there is a $\delta_\varepsilon > 0$ such that $\inf_{(h,g)\in\tilde{D}_\varepsilon} |\phi_P(h,g)| > \delta_\varepsilon$, since ϕ_P is continuous by Lemma C.12. Also, $\hat{\sigma}_{P_n}$ is uniformly bounded in (h,g) and ω, so there is a $\delta'_\varepsilon > 0$ such that for all $\omega \in \Omega$, $\inf_{(h,g)\in\tilde{D}_\varepsilon} |\phi_P(h,g) / (\xi_0 \vee \hat{\sigma}_{P_n}(h,g))| > \delta'_\varepsilon$. Thus if $\tilde{D}_\varepsilon \neq \varnothing$, we have

$$\begin{aligned}
&\lim_{n\to\infty} \mathbb{P}^*\left(\overrightarrow{d_H}\left(\widehat{\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}}, \Psi_{\bar{\mathcal{H}}\times\mathcal{G}}\right) > \varepsilon\right) \\
&= \lim_{n\to\infty} \mathbb{P}^*\left(\sup_{(h,g)\in\widehat{\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}}} \inf_{(h',g')\in\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}} \rho_P\left((h,g),(h',g')\right) > \varepsilon\right) \\
&\leq \lim_{n\to\infty} \mathbb{P}^*\left(\begin{array}{c} \sup_{(h,g)\in\widehat{\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}}\setminus\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}} \left|\frac{\phi_P(h,g)}{\xi_0\vee\hat{\sigma}_{P_n}(h,g)}\right| > \delta'_\varepsilon, \\ \sup_{(h,g)\in\widehat{\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}}\setminus\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}} \sqrt{T_n}\left|\frac{\hat{\phi}_{P_n}(h,g)}{\xi_0\vee\hat{\sigma}_{P_n}(h,g)}\right| \leq \tau_n \end{array}\right).
\end{aligned}$$

By Lemma C.11, we have that $\hat{\phi}_{P_n} \to \phi_P$ almost uniformly. Thus there is a measurable set A with $\mathbb{P}(A) \geq 1-\varepsilon$ such that for sufficiently large n,

$$\sup_{(h,g)\in\widehat{\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}}\setminus\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}} \left|\frac{\hat{\phi}_{P_n}(h,g)}{\xi_0\vee\hat{\sigma}_{P_n}(h,g)}\right| \geq \sup_{(h,g)\in\widehat{\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}}\setminus\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}} \left|\frac{\phi_P(h,g)}{\xi_0\vee\hat{\sigma}_{P_n}(h,g)}\right| - \frac{\delta'_\varepsilon}{2}$$

uniformly on A. Thus we now have that

$$\begin{aligned}
&\lim_{n\to\infty} \mathbb{P}^*\left(\overrightarrow{d_H}\left(\widehat{\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}}, \Psi_{\bar{\mathcal{H}}\times\mathcal{G}}\right) > \varepsilon\right) \\
&\leq \lim_{n\to\infty} \mathbb{P}^*\left(\begin{array}{c} \left\{\sup_{(h,g)\in\widehat{\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}}\setminus\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}} \left|\frac{\phi_P(h,g)}{\xi_0\vee\hat{\sigma}_{P_n}(h,g)}\right| > \delta'_\varepsilon\right\} \\ \cap\left\{\sup_{(h,g)\in\widehat{\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}}\setminus\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}} \sqrt{T_n}\left|\frac{\hat{\phi}_{P_n}(h,g)}{\xi_0\vee\hat{\sigma}_{P_n}(h,g)}\right| \leq \tau_n\right\}\cap A \end{array}\right) + \mathbb{P}(A^c) \\
&\leq \lim_{n\to\infty} \mathbb{P}^*\left(\sqrt{\frac{T_n}{n}}\frac{\delta'_\varepsilon}{2} < \sup_{(h,g)\in\widehat{\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}}\setminus\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}} \sqrt{\frac{T_n}{n}}\left|\frac{\hat{\phi}_{P_n}(h,g)}{\xi_0\vee\hat{\sigma}_{P_n}(h,g)}\right| \leq \frac{\tau_n}{\sqrt{n}}\right) + \varepsilon = \varepsilon,
\end{aligned}$$

because $\tau_n/\sqrt{n} \to 0$ as $n \to \infty$. Here, ε can be arbitrarily small. ■

Proof of Theorem 3.2. (i). Fix $\psi \in C\left(\Xi\times\bar{\mathcal{H}}\times\mathcal{G}\right)$ under the $\rho_{\xi hg}$ defined in (C.26). It is easy to show that $\Xi\times\bar{\mathcal{H}}\times\mathcal{G}$ is compact under $\rho_{\xi hg}$, and thus ψ is uniformly continuous on $\Xi\times\bar{\mathcal{H}}\times\mathcal{G}$. This implies that for every $\varepsilon > 0$, there is a $\delta > 0$ such that $|\psi(\xi',h',g') - \psi(\xi,h,g)| \leq \varepsilon/\nu(\Xi)$ for all $(\xi,h,g),(\xi',h',g') \in \Xi\times\bar{\mathcal{H}}\times\mathcal{G}$ with $\rho_{\xi hg}((\xi',h',g'),(\xi,h,g)) \leq \delta$. Also, by the constructions of $\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}$ in (25) and $\widehat{\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}}$ in (C.47), we have that

$$\begin{aligned}
&\left|\mathcal{I}\circ\mathcal{S}_{\widehat{\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}}}(\psi) - \mathcal{I}\circ\mathcal{S}_{\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}}(\psi)\right| \\
&\leq \nu(\Xi) \sup_{\rho_{\xi hg}((\xi',h',g'),(\xi,h,g))\leq d_H\left(\widehat{\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}},\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}\right)} \left|\psi\left(\xi',h',g'\right) - \psi(\xi,h,g)\right|.
\end{aligned}$$

By Lemma C.17, this implies that

$$\mathbb{P}^*\left(\left|\mathcal{I}\circ\mathcal{S}_{\widehat{\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}}}(\psi)-\mathcal{I}\circ\mathcal{S}_{\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}}(\psi)\right|>\varepsilon\right)\leq\mathbb{P}^*\left(d_H\left(\widehat{\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}},\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}\right)>\delta\right)\to 0.$$

Notice that

$$\left|\mathcal{I}\circ\mathcal{S}_{\widehat{\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}}}(\psi_1)-\mathcal{I}\circ\mathcal{S}_{\widehat{\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}}}(\psi_2)\right|\leq\nu(\Xi)\left\|\psi_1-\psi_2\right\|_\infty$$

for all $\psi_1,\psi_2\in\ell^\infty\left(\Xi\times\bar{\mathcal{H}}\times\mathcal{G}\right)$. By Lemma S.3.6 of Fang and Santos (2018), $\mathcal{I}\circ\mathcal{S}_{\widehat{\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}}}$ satisfies Assumption 4 of Fang and Santos (2018). Together with Lemma C.16, by repeating the proof of Theorem 3.2 of Fang and Santos (2018) with $\mathbb{G}_n^B=\sqrt{T_n^B}(\hat{\phi}_{P_n}^B-\hat{\phi}_{P_n})/\mathcal{M}(\hat{\sigma}_{P_n}^B)$, where $\mathbb{G}_n^B$ replaces $\mathbb{G}_n^*$ in their notation, we can show that

$$\sup_{f\in\mathrm{BL}_1(\mathbb{R})}\left|\begin{array}{c}E_W\left[f\left\{\mathcal{I}\circ\mathcal{S}_{\widehat{\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}}}\left(\frac{\sqrt{T_n^B}(\hat{\phi}_{P_n}^B-\hat{\phi}_{P_n})}{\mathcal{M}(\hat{\sigma}_{P_n}^B)}\right)\right\}\right]\\-E\left[f\left\{\mathcal{I}\circ\mathcal{S}_{\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}}\left(\frac{\mathbb{G}_0}{\mathcal{M}(\sigma_P)}\right)\right\}\right]\end{array}\right|\to 0 \qquad \text{(C.48)}$$

in outer probability, where $\mathbb{G}_0$ is the limit obtained in Lemma C.16 and $\mathbb{G}_0/\mathcal{M}(\sigma_P)$ is tight by Lemma C.16(i). Since the sample is finite, that is, we have only finitely many observations $\{(Y_i,D_i,Z_i)\}_{i=1}^n$ in the data set, by the constructions of $\widehat{\Psi_{\mathcal{H}\times\mathcal{G}}}$ in (27) and $\widehat{\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}}$ in (C.47) we have that

$$\mathcal{I}\circ\mathcal{S}_{\widehat{\Psi_{\mathcal{H}\times\mathcal{G}}}}\left(\frac{\sqrt{T_n^B}\left(\hat{\phi}_{P_n}^B-\hat{\phi}_{P_n}\right)}{\mathcal{M}\left(\hat{\sigma}_{P_n}^B\right)}\right)=\mathcal{I}\circ\mathcal{S}_{\widehat{\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}}}\left(\frac{\sqrt{T_n^B}\left(\hat{\phi}_{P_n}^B-\hat{\phi}_{P_n}\right)}{\mathcal{M}\left(\hat{\sigma}_{P_n}^B\right)}\right). \qquad \text{(C.49)}$$

Then (C.48) and (C.49) imply that

$$\sup_{f\in\mathrm{BL}_1(\mathbb{R})}\left|\begin{array}{c}E_W\left[f\left\{\mathcal{I}\circ\mathcal{S}_{\widehat{\Psi_{\mathcal{H}\times\mathcal{G}}}}\left(\frac{\sqrt{T_n^B}(\hat{\phi}_{P_n}^B-\hat{\phi}_{P_n})}{\mathcal{M}(\hat{\sigma}_{P_n}^B)}\right)\right\}\right]\\-E\left[f\left\{\mathcal{I}\circ\mathcal{S}_{\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}}\left(\frac{\mathbb{G}_0}{\mathcal{M}(\sigma_P)}\right)\right\}\right]\end{array}\right|\to 0 \qquad \text{(C.50)}$$

in outer probability. Let F denote the CDF of $\mathcal{I}\circ\mathcal{S}_{\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}}\left(\mathbb{G}_0/\mathcal{M}\left(\sigma_P\right)\right)$, and define $\hat{F}_n$ by

$$\hat{F}_n(c)=\mathbb{P}\left(\mathcal{I}\circ\mathcal{S}_{\widehat{\Psi_{\mathcal{H}\times\mathcal{G}}}}\left(\frac{\sqrt{T_n^B}\left(\hat{\phi}_{P_n}^B-\hat{\phi}_{P_n}\right)}{\mathcal{M}\left(\hat{\sigma}_{P_n}^B\right)}\right)\leq c\,\middle|\,\{(Y_i,D_i,Z_i)\}_{i=1}^\infty\right).$$[12]

Since by assumption F is continuous and increasing at $c_{1-\alpha}$, by a proof similar to that of Theorem S.1.1 of Fang and Santos (2018) together with (C.50) in this paper, we can

[12]This conditional probability given $\{(Y_i,D_i,Z_i)\}_{i=1}^\infty$ is numerically equal to that given $\{(Y_i,D_i,Z_i)\}_{i=1}^n$ in (31).

conclude that for each $\varepsilon > 0$,

$$\mathbb{P}^*(|\hat{c}_{1-\alpha} - c_{1-\alpha}| > \varepsilon) \to 0. \tag{C.51}$$

By the definitions of $\mathbb{G}$ (in the proof of Lemma 3.1) and $\mathbb{G}_0$ (in Lemma C.16), together with the linearity of $\mathcal{L}'_P$, we have that $\mathbb{G} = \mathbb{G}_0 + \Lambda(P)^{1/2}\mathcal{L}'_P(Q_0)$. Let $H_n = \sqrt{n}(P_n - P)$. By Lemma C.10, $\|H_n - Q_0\|_\infty \to 0$ as $n \to \infty$. Notice that $P_n = P + n^{-1/2}H_n$. By Lemma C.9, we have that

$$\begin{aligned}
&\lim_{n\to\infty} \sup_{(h,g)\in\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}} \left| \frac{\mathcal{L}(P_n)(h,g) - \mathcal{L}(P)(h,g)}{n^{-1/2}} - \mathcal{L}'_P(Q_0)(h,g) \right| \\
&\leq \lim_{n\to\infty} \sup_{(h,g)\in\bar{\mathcal{H}}\times\mathcal{G}} \left| \frac{\mathcal{L}\left(P + n^{-1/2}H_n\right)(h,g) - \mathcal{L}(P)(h,g)}{n^{-1/2}} - \mathcal{L}'_P(Q_0)(h,g) \right| = 0.
\end{aligned} \tag{C.52}$$

By construction, $\mathcal{L}(P) = 0$ on $\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}$ because $\mathcal{L}(P) = \phi_P$. By assumption, we have that $\mathcal{L}(P_n) = \phi_{P_n} \leq 0$ on $\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}$ and (C.52) implies that $\mathcal{L}'_P(Q_0) \leq 0$ on $\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}$. Thus we have that $\mathbb{G} \leq \mathbb{G}_0$ and $\mathcal{I} \circ \mathcal{S}_{\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}}(\mathbb{G}/\mathcal{M}(\sigma_P)) \leq \mathcal{I} \circ \mathcal{S}_{\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}}(\mathbb{G}_0/\mathcal{M}(\sigma_P))$. Since $\mathbb{G}/\mathcal{M}(\sigma_P) \in \ell^\infty\left(\Xi \times \bar{\mathcal{H}} \times \mathcal{G}\right)$, where $\ell^\infty\left(\Xi \times \bar{\mathcal{H}} \times \mathcal{G}\right)$ is a Banach space under $\|\cdot\|_\infty$ and $\mathbb{G}$ is tight by Lemma 3.1, we have that $\mathbb{G}/\mathcal{M}(\sigma_P)$ is tight (hence separable[13]) and is Radon by Theorem 7.1.7 of Bogachev (2007). Since $\mathcal{I} \circ \mathcal{S}_{\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}}$ is continuous and convex, Theorem 11.1(i) of Davydov et al. (1998) implies that the CDF of $\mathcal{I} \circ \mathcal{S}_{\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}}(\mathbb{G}/\mathcal{M}(\sigma_P))$ is everywhere continuous except possibly at the point

$$r_0 = \inf\left\{r : \mathbb{P}\left(\mathcal{I} \circ \mathcal{S}_{\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}}(\mathbb{G}/\mathcal{M}(\sigma_P)) \leq r\right) > 0\right\}.$$

Because $\mathcal{I} \circ \mathcal{S}_{\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}}(\mathbb{G}/\mathcal{M}(\sigma_P)) \leq \mathcal{I} \circ \mathcal{S}_{\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}}(\mathbb{G}_0/\mathcal{M}(\sigma_P))$, we have that

$$r_0 \leq \inf\left\{r : \mathbb{P}\left(\mathcal{I} \circ \mathcal{S}_{\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}}(\mathbb{G}_0/\mathcal{M}(\sigma_P)) \leq r\right) > 0\right\} < c_{1-\alpha},$$

where the last inequality follows from that the CDF of $\mathcal{I} \circ \mathcal{S}_{\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}}(\mathbb{G}_0/\mathcal{M}(\sigma_P))$ is continuous and increasing at $c_{1-\alpha}$. This implies that the CDF of $\mathcal{I} \circ \mathcal{S}_{\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}}(\mathbb{G}/\mathcal{M}(\sigma_P))$ is continuous at $c_{1-\alpha}$. Now by (26) and (C.51) in this paper, together with Example 1.4.7 (Slutsky's lemma), Theorem 1.3.6 (continuous mapping), and Theorem 1.3.4(vi) of van der Vaart and Wellner

[13]See the definition of separability in van der Vaart and Wellner (1996, p. 17). The closure of a separable subset of a metric space is separable.

(1996), we conclude that

$$\lim_{n\to\infty}\mathbb{P}^*\left(\sqrt{T_n}\mathcal{I}\circ\mathcal{S}\left(\frac{\hat{\phi}_{P_n}}{\mathcal{M}(\hat{\sigma}_{P_n})}\right)>\hat{c}_{1-\alpha}\right)=\mathbb{P}\left(\mathcal{I}\circ\mathcal{S}_{\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}}\left(\frac{\mathbb{G}}{\mathcal{M}(\sigma_P)}\right)>c_{1-\alpha}\right)\leq\alpha, \tag{C.53}$$

where the inequality follows from that $c_{1-\alpha}$ is the $1-\alpha$ quantile for $\mathcal{I}\circ\mathcal{S}_{\Psi_{\bar{\mathcal{H}}\times\mathcal{G}}}(\mathbb{G}_0/\mathcal{M}(\sigma_P))$. If, in addition, $P_n = P$ for all n, then by Assumption 3.2 we have that $v_0 = 0$ and hence $Q_0 = 0$. This implies that $\mathbb{G} = \mathbb{G}_0$ and that the inequality in (C.53) holds with equality.

(ii). Let $\hat{c}'_{1-\alpha}$ be the bootstrap critical value obtained using the bootstrap test statistic $\mathcal{I}\circ\mathcal{S}(\sqrt{T_n^B}(\hat{\phi}^B_{P_n}-\hat{\phi}_{P_n})/\mathcal{M}(\hat{\sigma}^B_{P_n}))$ in place of $\mathcal{I}\circ\mathcal{S}_{\widehat{\Psi_{\mathcal{H}\times\mathcal{G}}}}(\sqrt{T_n^B}(\hat{\phi}^B_{P_n}-\hat{\phi}_{P_n})/\mathcal{M}(\hat{\sigma}^B_{P_n}))$ in the test procedure in Section 3.1.1. By arguments similar to those in the proof of part (i), we can show that $\hat{c}'_{1-\alpha}\to c'_{1-\alpha}$ in outer probability, where $c'_{1-\alpha}$ is the $1-\alpha$ quantile for $\mathcal{I}\circ\mathcal{S}(\mathbb{G}_0/\mathcal{M}(\sigma_P))$.[14] Clearly, $\hat{c}'_{1-\alpha}\geq\hat{c}_{1-\alpha}$ by construction. By Lemma C.11, $\hat{\phi}_{P_n}/\mathcal{M}(\hat{\sigma}_{P_n})\to\phi_P/\mathcal{M}(\sigma_P)$ in $\ell^\infty(\Xi\times\bar{\mathcal{H}}\times\mathcal{G})$ almost uniformly, and hence almost uniformly

$$\mathcal{I}\circ\mathcal{S}_{\mathcal{H}\times\mathcal{G}}\left(\frac{\hat{\phi}_{P_n}}{\mathcal{M}(\hat{\sigma}_{P_n})}\right)\to\mathcal{I}\circ\mathcal{S}_{\mathcal{H}\times\mathcal{G}}\left(\frac{\phi_P}{\mathcal{M}(\sigma_P)}\right)>0,$$

where the inequality follows from the assumption that the H_0 in (15) is false with $Q = P$. Thus we have that $[\mathcal{I}\circ\mathcal{S}_{\mathcal{H}\times\mathcal{G}}(\sqrt{T_n}\hat{\phi}_{P_n}/\mathcal{M}(\hat{\sigma}_{P_n}))]^{-1}\to 0$ almost uniformly ($T_n/n\to\Lambda(P)$ almost uniformly by Lemma C.11). By Lemmas 1.9.3(ii) and 1.10.2(iii), Example 1.4.7 (Slutsky's lemma), and Theorems 1.3.6 (continuous mapping) and 1.3.4(vi) of van der Vaart and Wellner (1996), we now conclude that

$$\mathbb{P}^*\left(\mathcal{I}\circ\mathcal{S}_{\mathcal{H}\times\mathcal{G}}\left(\frac{\sqrt{T_n}\hat{\phi}_{P_n}}{\mathcal{M}(\hat{\sigma}_{P_n})}\right)>\hat{c}_{1-\alpha}\right)\geq\mathbb{P}^*\left(\mathcal{I}\circ\mathcal{S}_{\mathcal{H}\times\mathcal{G}}\left(\frac{\sqrt{T_n}\hat{\phi}_{P_n}}{\mathcal{M}(\hat{\sigma}_{P_n})}\right)>\hat{c}'_{1-\alpha}\right)\to 1.$$

■

D Monotonicity Condition with Unspecified Directions for Unordered Treatment

In Section 2.3, we mentioned that the test can be extended for the monotonicity condition with unspecified directions. We now show details for this extension. Define 2^J different

[14]Here, we implicitly assume that the CDF of $\mathcal{I}\circ\mathcal{S}(\mathbb{G}_0/\mathcal{M}(\sigma_P))$ is continuous and strictly increasing at $c'_{1-\alpha}$. Theorem 11.1 of Davydov et al. (1998) implies that the CDF of $\mathcal{I}\circ\mathcal{S}(\mathbb{G}_0/\mathcal{M}(\sigma_P))$ is differentiable and has a positive derivative everywhere except at countably many points in its support, provided that $\mathcal{I}\circ\mathcal{S}(\mathbb{G}_0/\mathcal{M}(\sigma_P))$ is not a constant. By construction, $\mathcal{I}\circ\mathcal{S}(\mathbb{G}_0/\mathcal{M}(\sigma_P))$ is not a constant in general cases.

J-dimensional binary vectors by $v_1, \ldots, v_{2^J}$ with

$$v_1 = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, v_2 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \ldots, v_{2^J} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}.$$

Let $\mathrm{L} : \mathcal{D} \to \{1, \ldots, J\}$ map $d \in \mathcal{D}$ to d's index in $\mathcal{D}$ such that if $d = d_j$, then $\mathrm{L}(d) = j$. For every $q \in \{1, \ldots, 2^J\}$, define a function $f_q : \mathcal{D} \to \{1, -1\}$ by $f_q(d) = (-1)^{v_q(\mathrm{L}(d))}$, where $v_q(j)$ denotes the jth element of v_q. If the instrument Z is valid for the unordered treatment D as defined in Assumption 2.4 with (iii) replaced by Assumption 2.3, then for all $z_j, z_k \in \mathcal{Z}$ with $j < k$, there is a $q \in \{1, \ldots, 2^J\}$ such that

$$f_q(d) \cdot \{\mathbb{P}(Y \in B, D = d | Z = z_j) - \mathbb{P}(Y \in B, D = d | Z = z_k)\} \leq 0$$

for every $d \in \mathcal{D}$ and every closed interval B. Then for every $q \in \{1, \ldots, 2^J\}$, we define

$$\mathcal{H}_q = \left\{ f_q(d) \cdot 1_{B \times \{d\} \times \mathbb{R}} : B \text{ is a closed interval in } \mathbb{R}, d \in \mathcal{D} \right\} \text{ and}$$
$$\bar{\mathcal{H}}_q = \left\{ f_q(d) \cdot 1_{B \times \{d\} \times \mathbb{R}} : B \text{ is a closed, open, or half-closed interval in } \mathbb{R}, d \in \mathcal{D} \right\}.$$

Also, define function spaces

$$\mathcal{H} = \cup_{q=1}^{2^J} \mathcal{H}_q, \bar{\mathcal{H}} = \cup_{q=1}^{2^J} \bar{\mathcal{H}}_q, \text{ and } \mathcal{G} = \left\{ \left(1_{\mathbb{R} \times \mathbb{R} \times \{z_j\}}, 1_{\mathbb{R} \times \mathbb{R} \times \{z_k\}} \right) : j, k \in \{1, \ldots, K\}, j < k \right\}. \tag{D.54}$$

Let ϕ_Q be defined as in (14) with $\mathcal{H}$ and $\mathcal{G}$ defined in (D.54). Now we obtain the testable implication for Assumption 2.4 with (iii) replaced by Assumption 2.3:

$$H_0 : \max_{g \in \mathcal{G}} \min_{q \in \{1, \ldots, 2^J\}} \sup_{h \in \mathcal{H}_q} \phi_Q(h, g) = 0, \tag{D.55}$$

if the underlying distribution of the data is Q. The test proposed in Section 3.3 can be generalized for the H_0 in (D.55).

E Additional Monte Carlo Studies

E.1 Degenerate Case under Null

In Section 3.1, we discussed the case where $\mathbb{T}_0 = 0$. In this section, we design a DGP such that $\mathbb{T}_0 = 0$ to show the performance of the test in this case. We let this DGP be the same as that designed in Section 4.1, except that we let $D_0 = 2 \times 1\{V \leq 0.328\} + 1\{0.328 < V \leq 0.658\}$, $D_1 = 2 \times 1\{V \leq 0.329\} + 1\{0.329 < V \leq 0.659\}$, and $D_2 = 2 \times 1\{V \leq 0.33\} + 1\{0.33 < V \leq 0.66\}$. Then it can be shown that $\mathbb{T}_0 = 0$ in this setting. We used $\eta = 0$ as discussed in Section 3.1. As suggested in Section 4.1, τ_n could be set to 2. Table 5 shows that the rejection rates are well controlled by the nominal significance level 0.05 in this case.

Table 5: Rejection Rates under H_0 ($\mathbb{T}_0 = 0$) for Multivalued D and Multivalued Z

τ_n	ξ for δ_ξ										$\bar{\nu}_\xi$
	0.07	0.1	0.13	0.16	0.19	0.22	0.25	0.28	0.3	1	
0.1	0.117	0.102	0.091	0.104	0.100	0.094	0.090	0.090	0.090	0.090	0.104
0.5	0.080	0.068	0.063	0.072	0.071	0.067	0.071	0.071	0.071	0.071	0.077
1	0.073	0.055	0.048	0.057	0.064	0.055	0.057	0.057	0.057	0.057	0.053
2	0.066	0.045	0.042	0.048	0.052	0.050	0.050	0.050	0.050	0.050	0.045
3	0.066	0.045	0.042	0.048	0.052	0.050	0.050	0.050	0.050	0.050	0.045
4	0.066	0.045	0.042	0.048	0.052	0.050	0.050	0.050	0.050	0.050	0.045
∞	0.066	0.045	0.042	0.048	0.052	0.050	0.050	0.050	0.050	0.050	0.045

E.2 Multivalued Treatment with Covariates

In Section 5, we used the data set of Card (1993) to illustrate the application of the proposed test in practice. We revisited this empirical example and reconducted the test with conditioning covariates added into the model. Due to the limitation on the computation power, we added two conditioning variables (the variables "south66" and "black") from the data set. When we chose the values of ξ, we employed the empirical variance formula in (21) to calculate an empirical bound for $\hat{\sigma}_{P_n}$. Specifically, we let $T_n = n \cdot \prod_{k=1}^{K} \prod_{l=1}^{L} \hat{P}_n(1_{\mathbb{R}\times\mathbb{R}\times\{z_k\}\times\{x_l\}})$ and used the first inequality in (22) to find the empirical bound. Table 6 shows the p-values obtained from the test. The p-values are lower than those in Section 5. One possible reason is that when conditioning covariates are included, the number of observations for each category (z_k, x_l) is small. Thus the p-values are different from those in the case where no conditioning covariates are included. But the results are consistent with those in Section 5 and show that the validity of the instrument is not rejected.

Table 6: p-values Obtained from the Proposed Test for Each Measure ν with Covariates

ξ for δ_ξ									$\bar{\nu}_\xi$
0.0001	0.00013	0.00016	0.00019	0.00022	0.00025	0.00028	0.00031	0.00034	
0.673	0.541	0.519	0.469	0.477	0.489	0.489	0.489	0.489	0.522

E.3 Unordered Treatment

In this section, we designed Monte Carlo simulations for the case where D is an unordered random variable with $D \in \{a, b, c\}$. For simplicity, we let $Z \in \{0, 1\}$. We also consider the presence of a conditioning covariate $X \in \{0, 1\}$. The measure ν was set to be a Dirac measure δ_ξ centered at one of the following values of ξ: 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, and 0.1, or to be a probability measure $\bar{\nu}_\xi$ that assigns equal probabilities (weights) to the values of ξ listed above. The nominal significance level α was set to 0.05. To expedite the simulation, we employed the warp-speed method of Giacomini et al. (2013).

E.3.1 Size Control and Tuning Parameter Selection

The first set of simulations was designed to investigate the size of the test and the selection of the tuning parameter. For this set of simulations, we set n to 2000 and τ_n to 0.1, 0.5, 1, 2, 3, 4, and ∞. We compared the rejection rates obtained using each of these values of τ_n and decided which value would be a good option for sample sizes close to 2000. The simulation consisted of 1000 Monte Carlo iterations and 1000 bootstrap iterations. We let $U \sim \mathrm{Unif}(0, 1)$, $U_X \sim \mathrm{Unif}(0, 1)$, $V \sim \mathrm{Unif}(0, 1)$, $N_a \sim \mathrm{N}(0, 1)$, $N_b \sim \mathrm{N}(1, 1)$, $N_c \sim \mathrm{N}(2, 1)$, $Z = 1\{U \leq 0.5\}$ ($\mathbb{P}(Z = 1) = 0.5$), $X = 1\{U_X \leq 0.5\}$,

$$D_z = \left\{ \begin{array}{c|c} a & V > 0.6 \\ b & 0.5 < V \leq 0.6 \\ c & 0 < V \leq 0.5 \end{array} \right.$$

for $z \in \{0, 1\}$, $D = D_z$ if $Z = z$ with $z \in \{0, 1\}$, and $Y = \sum_{d \in \{a,b,c\}} 1\{D = d\} \times N_d$. All the variables U, U_X, V, N_a, N_b, and N_c were set to be mutually independent. Assumption 2.4 holds in this case with $\mathcal{C} = \{(a, 0, 1), (b, 1, 0), (c, 1, 0)\}$.

Table 7 shows the results of the simulations. The rejection rates were influenced by the values of τ_n and ξ. For each measure ν, a smaller τ_n yields greater rejection rates by construction. For $\tau_n = 2$, all the rejection rates were close to those for $\tau_n = \infty$ (the conservative case). Similar to the pattern of the results shown in Kitagawa (2015) and Section 4.1, some rejection rates for $\tau_n = 2$ with δ_ξ centered at particular values of ξ were slightly upwardly biased compared to the nominal size. Overall, however, the results showed good performance of the test in terms of size control. When sample sizes are less than or close to 2000, we suggest using $\tau_n = 2$ in practice to achieve good size control without a significant power loss. When the sample size increases, τ_n should be increased accordingly. It is also worth noting that when we used the measure $\bar{\nu}_\xi$, the rejection rates could be well controlled by the nominal significance level. Thus if we have no additional information about the choice of ξ, $\bar{\nu}_\xi$ can be a default choice for us.

Table 7: Rejection Rates under H_0 for Unordered D

τ_n	ξ for δ_ξ										$\bar{\nu}_\xi$
	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1	
0.1	0.137	0.137	0.118	0.102	0.111	0.104	0.095	0.120	0.116	0.116	0.136
0.5	0.092	0.093	0.076	0.082	0.061	0.069	0.072	0.084	0.075	0.075	0.082
1	0.057	0.070	0.065	0.067	0.059	0.065	0.065	0.055	0.052	0.052	0.069
2	0.009	0.055	0.056	0.061	0.058	0.064	0.058	0.045	0.049	0.049	0.053
3	0.006	0.050	0.054	0.061	0.058	0.064	0.058	0.045	0.049	0.049	0.053
4	0.006	0.050	0.054	0.061	0.058	0.064	0.058	0.045	0.049	0.049	0.053
∞	0.006	0.050	0.054	0.061	0.058	0.064	0.058	0.045	0.049	0.049	0.053

E.3.2 Rejection Rates against Fixed Alternatives

The second set of simulations was designed to investigate the power of the test. A total of five DGPs were considered. Sample sizes were set to $n = 200$, 600, 1000, 1100, and 2000. The probability $\mathbb{P}(Z = 1) = r_n$, with $r_n = 1/2$, $1/6$, $1/2$, $1/11$, and $1/2$ for the corresponding sample sizes. We set τ_n to 2, as suggested in the preceding set of simulations. Each simulation consisted of 500 Monte Carlo iterations and 500 bootstrap iterations. We let $U \sim \text{Unif}(0,1)$, $U_X \sim \text{Unif}(0,1)$, $V \sim \text{Unif}(0,1)$, $W \sim \text{Unif}(0,1)$, $Z = 1\{U \leq r_n\}$, and $X = 1\{U_X \leq 0.5\}$. For DGPs (1)–(4), we let

$$D_z = \left\{ \begin{array}{c|c} a & V > 0.6 \\ b & 0.5 < V \leq 0.6 \\ c & 0 < V \leq 0.5 \end{array} \right.$$

for $z \in \{0,1\}$, $D = D_z$ if $Z = z$ with $z \in \{0,1\}$, $N_Z \sim \text{N}(0,1)$, $N_{az} = N_Z$ for $z \in \{0,1\}$, $N_{bz} = N_Z$ for $z \in \{0,1\}$, and $N_{c1} = N_Z$.

(1): $N_{c0} \sim \text{N}(-0.7, 1)$ and $Y = \sum_{z=0}^{1} 1\{Z = z\} \times (\sum_{d\in\{a,b,c\}} 1\{D = d\} \times N_{dz})$.

(2): $N_{c0} \sim \text{N}(0, 1.675^2)$ and $Y = \sum_{z=0}^{1} 1\{Z = z\} \times (\sum_{d\in\{a,b,c\}} 1\{D = d\} \times N_{dz})$.

(3): $N_{c0} \sim \text{N}(0, 0.515^2)$ and $Y = \sum_{z=0}^{1} 1\{Z = z\} \times (\sum_{d\in\{a,b,c\}} 1\{D = d\} \times N_{dz})$.

(4): $N_{c0a} \sim \text{N}(-1, 0.125^2)$, $N_{c0b} \sim \text{N}(-0.5, 0.125^2)$, $N_{c0c} \sim \text{N}(0, 0.125^2)$, $N_{c0d} \sim \text{N}(0.5, 0.125^2)$, $N_{c0e} \sim \text{N}(1, 0.125^2)$, $N_{c0} = 1\{W \leq 0.15\} \times N_{c0a} + 1\{0.15 < W \leq 0.35\} \times N_{c0b} + 1\{0.35 < W \leq 0.65\} \times N_{c0c} + 1\{0.65 < W \leq 0.85\} \times N_{c0d} + 1\{W > 0.85\} \times N_{c0e}$, and $Y = \sum_{z=0}^{1} 1\{Z = z\} \times (\sum_{d\in\{a,b,c\}} 1\{D = d\} \times N_{dz})$.

For DGP (5), we let

$$D_0 = \left\{ \begin{array}{c|c} a & V > 0.6 \\ b & 0.5 < V \leq 0.6 \\ c & 0 < V \leq 0.5 \end{array} \right., D_1 = \left\{ \begin{array}{c|c} a & V > 0.3 \\ b & 0.2 < V \leq 0.3 \\ c & 0 < V \leq 0.2 \end{array} \right..$$

(5): Let $N_a \sim \mathrm{N}(0,1)$, $N_b \sim \mathrm{N}(1,1)$, $N_c \sim \mathrm{N}(2,1)$, $D = D_z$ if $Z = z$ with $z \in \{0,1\}$, and $Y = \sum_{d\in\{a,b,c\}} 1\{D = d\} \times N_d$.

All the variables U, U_X, V, N_Z, N_{c0}, N_a, N_b, and N_c were set to be mutually independent.

Table 8 shows the rejection rates under DGPs (1)–(5), that is, the power of the test. For each DGP and each measure ν, the rejection rate increased as the sample size n was increased. The results for $\nu = \bar{\nu}_\xi$ showed that if we have no information about the choice of ξ, using the weighted average of the statistics over ξ is a desirable option. When $n > 200$, the rejection rates for using $\nu = \bar{\nu}_\xi$ were at a relatively high level compared to the results for using a Dirac measure.

Table 8: Rejection Rates under H_1 for Unordered D

DGP	n	ξ for δ_ξ										$\bar{\nu}_\xi$
		0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1	
(1)	200	0.000	0.090	0.188	0.256	0.324	0.336	0.326	0.290	0.306	0.306	0.222
	600	0.032	0.402	0.528	0.562	0.546	0.502	0.432	0.432	0.432	0.432	0.464
	1000	0.604	0.932	0.954	0.976	0.984	0.984	0.972	0.966	0.962	0.962	0.986
	1100	0.488	0.594	0.626	0.566	0.470	0.448	0.448	0.448	0.448	0.448	0.626
	2000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
(2)	200	0.000	0.006	0.044	0.112	0.134	0.096	0.060	0.050	0.044	0.044	0.034
	600	0.002	0.174	0.092	0.048	0.022	0.028	0.030	0.030	0.030	0.030	0.042
	1000	0.190	0.624	0.772	0.722	0.572	0.358	0.150	0.124	0.108	0.108	0.512
	1100	0.236	0.074	0.048	0.036	0.044	0.042	0.042	0.042	0.042	0.042	0.078
	2000	0.992	0.998	0.998	0.998	0.970	0.898	0.642	0.456	0.398	0.398	0.976
(3)	200	0.000	0.160	0.334	0.398	0.452	0.460	0.494	0.484	0.490	0.490	0.364
	600	0.042	0.560	0.666	0.786	0.812	0.798	0.750	0.750	0.750	0.750	0.728
	1000	0.728	0.926	0.948	0.958	0.980	0.986	0.990	0.992	0.990	0.990	0.988
	1100	0.596	0.720	0.824	0.860	0.792	0.764	0.764	0.764	0.764	0.764	0.826
	2000	0.996	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
(4)	200	0.000	0.042	0.110	0.150	0.172	0.200	0.214	0.214	0.208	0.208	0.146
	600	0.026	0.326	0.382	0.396	0.428	0.442	0.414	0.404	0.404	0.404	0.436
	1000	0.210	0.472	0.572	0.576	0.618	0.702	0.706	0.746	0.774	0.774	0.704
	1100	0.326	0.444	0.530	0.568	0.504	0.444	0.444	0.444	0.444	0.444	0.580
	2000	0.790	0.930	0.948	0.954	0.962	0.956	0.968	0.978	0.982	0.982	0.986
(5)	200	0.162	0.900	0.958	0.968	0.974	0.974	0.984	0.988	0.988	0.988	0.970
	600	0.688	0.988	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	1100	0.974	1.000	1.000	1.000	1.000	0.996	0.996	0.996	0.996	0.996	1.000
	2000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

E.4 Comparison in Binary Case

The Monte Carlo experiments discussed in this section followed the design of Kitagawa (2015), where the treatment and the instrument were both binary, with $D \in \{0,1\}$ and $Z \in \{0,1\}$, and we compared our results with theirs. We simulated the limiting rejection rates using the approach proposed in the present paper and that proposed by Kitagawa (2015) with the same randomly generated data. In this special case, if the measure ν is set to be a Dirac measure, the asymptotic distribution of the test statistic under null can be written as $\sup_{f\in\mathcal{F}_b^*} \mathbb{G}_H(f)/(\xi \vee \sigma_H(f))$ in (32). We followed the discussion in Section 3.2 to construct the bootstrap critical value. That is, we approximated $\mathbb{G}_H$ and σ_H by $\mathbb{G}_H^B$ and σ_H^B following the bootstrap method of Kitagawa (2015). Then we estimated $\mathcal{F}_b^*$ by $\widehat{\mathcal{F}_b^*}$ in a way similar to (27), which is the key difference between our approach and that of Kitagawa (2015). Last, we constructed the bootstrap test statistic by $\sup_{f\in\widehat{\mathcal{F}_b^*}} \mathbb{G}_H^B(f)/(\xi \vee \sigma_H^B(f))$ and used it to create the critical value. Because of $\widehat{\mathcal{F}_b^*}$, our bootstrap test statistic can approximate the null distribution consistently and the power of the test can be improved. This new bootstrap test statistic is asymptotically equivalent to that in (30) under null, and the new critical value is asymptotically equivalent to $\hat{c}_{1-\alpha}$ in Section 3.1.1 under null.

Each simulation consisted of 1000 Monte Carlo iterations and 1000 bootstrap iterations. For each DGP, the measure ν was set to be a Dirac measure centered at $\xi = 0.07$, 0.22, 0.3, and 1. The nominal significance level α was set to 0.05.

E.4.1 Size Control and Tuning Parameter Selection

We first ran simulations to investigate the size of the test and the selection of the tuning parameter. As suggested in Section 4, for sample sizes less than 3000, we can use $\tau_n = 2$ for the tuning parameter. In this set of simulations, we set $n = 2000$ and $\tau_n = 1, 2, 3, 4, \infty$. For the DGP, we used $U \sim \text{Unif}(0,1)$, $V \sim \text{Unif}(0,1)$, $N_0 \sim \text{N}(0,1)$, $N_1 \sim \text{N}(1,1)$, $Z = 1\{U \le 0.5\}$, $D_0 = 1\{V \le 0.5\}$, $D_1 = 1\{V \le 0.5\}$, $D = \sum_{z=0}^{1} 1\{Z = z\} \times D_z$, and $Y = \sum_{d=0}^{1} 1\{D = d\} \times N_d$, where U, V, N_0, and N_1 were mutually independent. This DGP is equivalent to that used by Kitagawa (2015) to show the size control of their test. The results in Table 9 confirmed the conclusion from Table 1: For $\tau_n = 2$, the rejection rates were close to those for $\tau_n = \infty$ and close to the nominal size. Recall that a smaller tuning parameter τ_n yields greater power for the test. Thus we kept using $\tau_n = 2$ in this case.

E.4.2 Power Comparison

Four DGPs were considered for the power comparisons. The sample sizes were set to $n = 200$, 600, 1000, 1100, and 2000, and the tuning parameter was set to $\tau_n = 2$. The probability

Table 9: Rejection Rates under H_0 for Binary D and Binary Z

τ_n	ξ			
	0.07	0.22	0.3	1
1	0.077	0.052	0.048	0.069
2	0.058	0.048	0.040	0.067
3	0.056	0.046	0.040	0.067
4	0.056	0.046	0.040	0.067
∞	0.056	0.046	0.040	0.067

$\mathbb{P}(Z = 1) = r_n$ with $r_n = 1/2$, $1/6$, $1/2$, $1/11$, and $1/2$ for the corresponding sample sizes. We let $U \sim \text{Unif}(0, 1)$, $V \sim \text{Unif}(0, 1)$, $W \sim \text{Unif}(0, 1)$, $Z = 1\{U \leq r_n\}$, $D_0 = 1\{V \leq 0.45\}$, $D_1 = 1\{V \leq 0.55\}$, $D = \sum_{z=0}^{1} 1\{Z = z\} \times D_z$, $N_{00} \sim \text{N}(0, 1)$, $N_{01} \sim \text{N}(0, 1)$, and $N_{11} \sim \text{N}(0, 1)$.

(1): $N_{10} \sim \text{N}(-0.7, 1)$ and $Y = \sum_{z=0}^{1} 1\{Z = z\} \times (\sum_{d=0}^{1} 1\{D = d\} \times N_{dz})$.

(2): $N_{10} \sim \text{N}(0, 1.675^2)$ and $Y = \sum_{z=0}^{1} 1\{Z = z\} \times (\sum_{d=0}^{1} 1\{D = d\} \times N_{dz})$.

(3): $N_{10} \sim \text{N}(0, 0.515^2)$ and $Y = \sum_{z=0}^{1} 1\{Z = z\} \times (\sum_{d=0}^{1} 1\{D = d\} \times N_{dz})$.

(4): $N_{10a} \sim \text{N}(-1, 0.125^2)$, $N_{10b} \sim \text{N}(-0.5, 0.125^2)$, $N_{10c} \sim \text{N}(0, 0.125^2)$, $N_{10d} \sim \text{N}(0.5, 0.125^2)$, $N_{10e} \sim \text{N}(1, 0.125^2)$, $N_{10} = 1\{W \leq 0.15\} \times N_{10a} + 1\{0.15 < W \leq 0.35\} \times N_{10b} + 1\{0.35 < W \leq 0.65\} \times N_{10c} + 1\{0.65 < W \leq 0.85\} \times N_{10d} + 1\{W > 0.85\} \times N_{10e}$, and $Y = \sum_{z=0}^{1} 1\{Z = z\} \times (\sum_{d=0}^{1} 1\{D = d\} \times N_{dz})$.

All the variables U, V, N_{00}, N_{10}, N_{01}, and N_{11} were set to be mutually independent for each DGP. Table 10 shows a comparison of the powers of the two tests. The results suggest that the proposed test achieves a manifest power improvement over that of Kitagawa (2015).

E.4.3 Comparison in Empirical Application

We now revisit the empirical application in Section 5 and show the size and power comparisons with the test of Kitagawa (2015) using the data set of Card (1993). We follow Kitagawa (2015) and define T by $T = 1\{D \geq 16\}$. As discussed in Section 5, the instrument may not be valid for this coarsened treatment T. This has been verified by the empirical study of Kitagawa (2015). As shown in Table I of Kitagawa (2015), the sample size was sufficiently large (over 3000), so the null hypothesis was rejected with p-values exactly equal to 0 when no conditioning covariates were included in the model. To show the comparison of the proposed test and the test of Kitagawa (2015) in this empirical example, we randomly drew relatively small subsamples of sizes 700, 900, 1100, 1300, 1500, and 2000

Table 10: Rejection Rates under H_1 for Binary D and Binary Z

DGP	n	The Proposed Test				Test of Kitagawa (2015)			
		ξ				ξ			
		0.07	0.22	0.3	1	0.07	0.22	0.3	1
(1)	200	0.202	0.198	0.186	0.110	0.198	0.193	0.182	0.106
	600	0.300	0.434	0.418	0.180	0.240	0.406	0.375	0.144
	1000	0.874	0.915	0.919	0.804	0.855	0.883	0.894	0.714
	1100	0.309	0.493	0.452	0.163	0.263	0.451	0.423	0.153
	2000	0.997	0.999	1.000	0.997	0.996	0.999	0.999	0.993
(2)	200	0.105	0.095	0.059	0.004	0.090	0.084	0.046	0.003
	600	0.261	0.141	0.045	0.000	0.242	0.100	0.026	0.000
	1000	0.907	0.814	0.500	0.105	0.887	0.781	0.421	0.030
	1100	0.255	0.129	0.037	0.001	0.224	0.082	0.022	0.001
	2000	1.000	0.996	0.949	0.674	1.000	0.994	0.909	0.252
(3)	200	0.211	0.209	0.202	0.211	0.185	0.188	0.195	0.205
	600	0.203	0.427	0.473	0.351	0.191	0.377	0.458	0.331
	1000	0.664	0.769	0.816	0.831	0.654	0.739	0.785	0.796
	1100	0.229	0.442	0.487	0.341	0.203	0.399	0.443	0.321
	2000	0.950	0.982	0.992	0.995	0.949	0.971	0.987	0.992
(4)	200	0.080	0.082	0.073	0.036	0.079	0.082	0.073	0.036
	600	0.134	0.117	0.103	0.060	0.123	0.111	0.102	0.058
	1000	0.307	0.306	0.224	0.127	0.307	0.281	0.212	0.116
	1100	0.146	0.115	0.112	0.031	0.136	0.115	0.093	0.027
	2000	0.660	0.703	0.556	0.325	0.649	0.673	0.505	0.271

out of the full data set, and computed the empirical sizes and powers of the two tests using the same subsamples.

To compare the sizes, given the subsample $\{(Y_i, D_i)\}_{i=1}^m$ drawn randomly from the full data set with $m \in \{700, 900, 1100, 1300, 1500, 2000\}$, we let $Z_i = 0$ for $i = 1, \ldots, m/2$ and $Z_i = 1$ for $i = m/2 + 1, \ldots, m$. Then we used the sample $\{(Y_i, D_i, Z_i)\}_{i=1}^m$ to compute the sizes of the two tests. As shown in Table 11, the rejection rates of the proposed test under null are slightly higher than those of the test of Kitagawa (2015). Both are close to the nominal significance level 0.05.

We used the subsample $\{(Y_i, D_i, Z_i)\}_{i=1}^m$ drawn randomly from the full data set to compute the powers of the two tests. As shown in Table 12, the proposed test achieves a manifest power improvement over that of Kitagawa (2015) when the samples are relatively small.

Table 11: Empirical Sizes of the Two Tests for Each ξ in Empirical Application

n	ξ for δ_ξ									
	0.07	0.1	0.13	0.16	0.19	0.22	0.25	0.28	0.3	1
700	0.071	0.071	0.071	0.076	0.068	0.064	0.061	0.055	0.061	0.056
900	0.066	0.066	0.065	0.064	0.065	0.068	0.056	0.058	0.051	0.064
1100	0.072	0.072	0.070	0.059	0.055	0.059	0.054	0.051	0.053	0.048
1300	0.064	0.064	0.061	0.062	0.069	0.074	0.075	0.071	0.065	0.060
1500	0.045	0.047	0.046	0.046	0.045	0.039	0.050	0.049	0.066	0.062
2000	0.049	0.050	0.049	0.050	0.048	0.051	0.053	0.045	0.046	0.038

(a) Empirical Sizes of the Test of Kitagawa (2015)

n	ξ for δ_ξ									
	0.07	0.1	0.13	0.16	0.19	0.22	0.25	0.28	0.3	1
700	0.075	0.075	0.075	0.078	0.068	0.067	0.061	0.059	0.062	0.056
900	0.066	0.066	0.066	0.064	0.065	0.068	0.056	0.059	0.051	0.064
1100	0.072	0.072	0.070	0.060	0.061	0.060	0.055	0.054	0.053	0.048
1300	0.066	0.066	0.062	0.065	0.069	0.080	0.079	0.071	0.065	0.060
1500	0.045	0.047	0.046	0.046	0.045	0.040	0.050	0.049	0.066	0.062
2000	0.050	0.050	0.051	0.050	0.048	0.051	0.053	0.045	0.046	0.038

(b) Empirical Sizes of the Proposed Test

E.5 Choices of Ξ and ν in Applications

As shown in the discussion for (23) and also in the discussion in Kitagawa (2015), ξ plays a role of bounding $\hat{\sigma}_{P_n}$ sufficiently away from zero. We set the support of ξ, Ξ, to be a closed subset of $[0, 1]$ such that $0 \notin \Xi$. Though our theoretical results show that all such Ξ and measures ν that satisfy Assumption 3.3 yield good asymptotic properties of the test, the choices of Ξ and ν may affect the finite sample performance of the test. For finite samples, if ξ is small (not far enough from 0), larger sample sizes would be needed for the test to achieve better size and power properties. In this section, we provide more simulation results for the choices of Ξ and ν. We then provide an empirical approach for choosing Ξ and ν in practice.

We followed the same constructions of simulations under H_0 as those in Section 4.1, and we set $\Xi = \{0.01, 0.02, 0.03, 0.04, 0.07, 0.1, 0.13, 0.16, 0.19, 0.22, 0.25, 1\}$. We set the sample sizes to 1000, 2000, and 3000. In this way, we investigate how small values of ξ would affect the finite sample performance of the test for different sample sizes. The measure ν was set to be a Dirac measure δ_ξ centered at each value in Ξ. The bootstrap iteration was set to 1000. We focus on the results for $\tau_n = 2$ which is chosen in Section 4. As shown in Table 13, for $\xi \geq 0.04$ and $n = 3000$, all rejection rates are close to the nominal significance

Table 12: Empirical Powers of the Two Tests for Each ξ in Empirical Application

n	ξ for δ_ξ									
	0.07	0.1	0.13	0.16	0.19	0.22	0.25	0.28	0.3	1
700	0.321	0.321	0.340	0.379	0.390	0.409	0.416	0.408	0.423	0.476
900	0.395	0.390	0.427	0.457	0.478	0.494	0.517	0.538	0.535	0.578
1100	0.534	0.557	0.584	0.633	0.627	0.657	0.659	0.682	0.699	0.739
1300	0.666	0.666	0.706	0.728	0.762	0.768	0.798	0.791	0.793	0.805
1500	0.742	0.779	0.800	0.805	0.808	0.808	0.831	0.844	0.836	0.864
2000	0.902	0.907	0.920	0.904	0.922	0.924	0.929	0.927	0.930	0.952

(a) Empirical Powers of the Test of Kitagawa (2015)

n	ξ for δ_ξ									
	0.07	0.1	0.13	0.16	0.19	0.22	0.25	0.28	0.3	1
700	0.336	0.336	0.344	0.408	0.417	0.424	0.446	0.476	0.482	0.511
900	0.404	0.411	0.445	0.512	0.513	0.537	0.546	0.584	0.584	0.654
1100	0.558	0.572	0.615	0.647	0.654	0.668	0.686	0.726	0.751	0.803
1300	0.674	0.725	0.756	0.796	0.803	0.828	0.847	0.854	0.865	0.858
1500	0.781	0.825	0.852	0.857	0.836	0.850	0.860	0.871	0.875	0.905
2000	0.910	0.926	0.942	0.935	0.935	0.940	0.947	0.945	0.951	0.975

(b) Empirical Powers of the Proposed Test

level $\alpha = 0.05$. (The rejection rates for some ξ are slightly upward biased. In application-based simulations, these rejection rates get close to α.) For small $\xi \in \{0.01, 0.02, 0.03\}$, most of the rejection rates are lower than α, but they increase as n increases. For example, the rejection rates for $\xi = 0.03$ are 0.014, 0.033, and 0.070 for $n = 1000$, 2000, and 3000, respectively. The rejection rates for $\xi = 0.02$ are 0.001, 0.006, and 0.011 for $n = 1000$, 2000, and 3000, respectively. For $\xi = 0.01$, the rejection rates are all 0. As discussed above, ξ is used to bound $\hat{\sigma}_{P_n}$ away from 0. When ξ is close to 0, the test may be conservative in finite samples. As the sample size increases, the rejection rates would converge to the nominal significance level, as shown for $\xi = 0.02$ and 0.03. We expect that when n gets larger, the rejection rate for $\xi = 0.01$ would converge to α.

E.5.1 Application-based Simulations for Choosing Ξ and ν

Since small values of ξ may affect the finite sample performance of the test, we introduce an empirical way of choosing Ξ and ν in finite samples. In practice, we suggest setting Ξ to be a (large) finite set of values and ν to be a Dirac measure centered at each value of Ξ or a probability measure that assigns equal weights to each value in Ξ. The results in Section 4 show that these choices work well in simulations. Recall that (20) and (22)

Table 13: Rejection Rates under H_0 for Small ξ

τ_n	ξ for δ_ξ $(n = 1000)$											
	0.01	0.02	0.03	0.04	0.07	0.1	0.13	0.16	0.19	0.22	0.25	1
0.1	0.174	0.174	0.174	0.174	0.200	0.179	0.173	0.137	0.119	0.102	0.104	0.104
0.5	0.097	0.097	0.097	0.105	0.136	0.116	0.115	0.089	0.071	0.069	0.070	0.070
1	0.004	0.027	0.037	0.070	0.108	0.084	0.076	0.072	0.065	0.063	0.063	0.063
2	0.000	0.001	0.014	0.052	0.105	0.077	0.063	0.069	0.055	0.061	0.058	0.058
3	0.000	0.001	0.012	0.052	0.105	0.077	0.062	0.069	0.055	0.061	0.058	0.058
4	0.000	0.001	0.011	0.052	0.105	0.077	0.062	0.069	0.055	0.061	0.058	0.058
∞	0.000	0.001	0.011	0.052	0.105	0.077	0.062	0.069	0.055	0.061	0.058	0.058
τ_n	ξ for δ_ξ $(n = 2000)$											
	0.01	0.02	0.03	0.04	0.07	0.1	0.13	0.16	0.19	0.22	0.25	1
0.1	0.159	0.159	0.159	0.167	0.150	0.121	0.120	0.105	0.103	0.112	0.119	0.119
0.5	0.086	0.086	0.087	0.084	0.079	0.076	0.089	0.061	0.069	0.064	0.058	0.058
1	0.014	0.023	0.052	0.066	0.071	0.061	0.075	0.051	0.055	0.047	0.050	0.050
2	0.000	0.006	0.033	0.053	0.058	0.054	0.065	0.047	0.049	0.036	0.036	0.036
3	0.000	0.006	0.033	0.053	0.056	0.054	0.064	0.047	0.049	0.036	0.033	0.033
4	0.000	0.006	0.033	0.053	0.056	0.054	0.064	0.047	0.048	0.035	0.032	0.032
∞	0.000	0.006	0.033	0.053	0.056	0.054	0.064	0.047	0.048	0.035	0.032	0.032
τ_n	ξ for δ_ξ $(n = 3000)$											
	0.01	0.02	0.03	0.04	0.07	0.1	0.13	0.16	0.19	0.22	0.25	1
0.1	0.198	0.198	0.194	0.185	0.122	0.108	0.096	0.096	0.108	0.092	0.092	0.092
0.5	0.074	0.074	0.112	0.122	0.092	0.070	0.068	0.074	0.064	0.069	0.069	0.069
1	0.017	0.023	0.077	0.089	0.079	0.060	0.047	0.068	0.056	0.058	0.061	0.061
2	0.000	0.011	0.070	0.083	0.073	0.050	0.037	0.050	0.050	0.055	0.048	0.048
3	0.000	0.011	0.055	0.083	0.073	0.048	0.037	0.050	0.050	0.049	0.048	0.048
4	0.000	0.011	0.055	0.083	0.073	0.048	0.037	0.050	0.050	0.049	0.048	0.048
∞	0.000	0.011	0.055	0.083	0.073	0.048	0.037	0.050	0.050	0.049	0.048	0.048

provide bounds for σ_P and $\hat{\sigma}_{P_n}$. For every finite sample, we can use positive values not larger than $\{1/2 \cdot (K-1)^{-(K-1)}\}^{1/2}$ to construct Ξ. Clearly, $1/2 \cdot (K-1)^{-(K-1)} < 1$ for all K. Thus, we can just include $\xi = 1$ in Ξ which leads to the unweighted KS test statistic, and we set the other values of ξ to be smaller than $\{1/2 \cdot (K-1)^{-(K-1)}\}^{1/2}$. To be more precise, we can calculate the bound $1/2 \cdot \max_{(g_1', g_2') \in \mathcal{G}} \{(T_n/n)/\hat{P}_n(g_2') + (T_n/n)/\hat{P}_n(g_1')\}^{1/2}$ and only include values smaller than this bound other than 1.

We revisit the application in Section 5 and show how to choose Ξ in practice. In this empirical example, $Z \in \{0, 1\}$, and it follows that $\hat{\sigma}_{P_n} \leq 1/2$, which was also mentioned in Kitagawa (2015). We first set Ξ to be a large finite set with

$$\Xi = \{0.01, 0.02, 0.03, 0.04, 0.07, 0.1, 0.13, 0.16, 0.19, 0.22, 0.25, 0.28, 0.3, 1\}.$$

Let $n_0 = \sum_{i=1}^n 1\{Z_i = 0\}$ and $n_1 = \sum_{i=1}^n 1\{Z_i = 1\}$. The following is the procedure for choosing Ξ for the finite sample $\{(Y_i, D_i, Z_i)\}_{i=1}^n$:

(1) Find the subsample of $\{(Y_i, D_i, Z_i)\}_{i=1}^n$ with $Z_i = 0$, and denote this subsample by

$\{(Y_i^0, D_i^0, Z_i^0)\}_{i=1}^{n_0}$.

(2) Randomly draw two samples from $\{(Y_i^0, D_i^0)\}_{i=1}^{n_0}$, denoted by $\{(Y_i^{00}, D_i^{00})\}_{i=1}^{n_0}$ and $\{(Y_i^{01}, D_i^{01})\}_{i=1}^{n_1}$.

(3) Let $Z_i^{00} = 0$ for all $i \in \{1, \ldots, n_0\}$ and $Z_i^{01} = 1$ for all $i \in \{1, \ldots, n_1\}$.

(4) Combine the two samples $\{(Y_i^{00}, D_i^{00}, Z_i^{00})\}_{i=1}^{n_0}$ and $\{(Y_i^{01}, D_i^{01}, Z_i^{01})\}_{i=1}^{n_1}$. Denote the combined sample by $\{(Y_i^c, D_i^c, Z_i^c)\}_{i=1}^{n}$.

(5) Compute the test statistic and the bootstrap critical value based on the sample $\{(Y_i^c, D_i^c, Z_i^c)\}_{i=1}^{n}$ and record the test results for Ξ.

(6) Repeat steps (2)–(5) many times and find the values in Ξ such that the corresponding rejection rates are close to α.

(7) Repeat steps (2)–(6) using the subsample of $\{(Y_i, D_i, Z_i)\}_{i=1}^{n}$ with $Z_i = 1$. Find the values in Ξ such that the corresponding rejection rates are close to α.

(8) The intersection of the two sets of values in Ξ from the above steps can be used in the application.

Table 14 shows the simulation results following the above procedure. Based on these simulation results, we suggest using $\Xi = \{0.03, 0.04, 0.07, 0.1, 0.13, 0.16, 0.19, 0.22, 0.3, 1\}$ for this application. We reconducted the test in Section 5 using this new Ξ. Table 15 shows that the test results are similar to those in Table 4.

Table 14: Application-based Rejection Rates under H_0 for Different ξ

τ_n	ξ for δ_ξ $(Z=0)$													
	0.01	0.02	0.03	0.04	0.07	0.1	0.13	0.16	0.19	0.22	0.25	0.28	0.3	1
0.1	0.180	0.180	0.180	0.181	0.163	0.153	0.160	0.178	0.176	0.206	0.228	0.243	0.250	0.291
0.5	0.096	0.096	0.096	0.095	0.087	0.087	0.084	0.084	0.086	0.087	0.102	0.099	0.099	0.090
1	0.037	0.055	0.062	0.068	0.063	0.065	0.058	0.058	0.066	0.066	0.065	0.063	0.065	0.067
2	0.005	0.034	0.041	0.048	0.053	0.056	0.054	0.043	0.049	0.057	0.052	0.048	0.048	0.053
3	0.000	0.020	0.035	0.043	0.053	0.056	0.054	0.043	0.049	0.057	0.052	0.048	0.048	0.052
4	0.000	0.017	0.032	0.043	0.053	0.056	0.054	0.043	0.049	0.057	0.052	0.048	0.048	0.052
∞	0.000	0.017	0.032	0.043	0.053	0.056	0.054	0.043	0.049	0.057	0.052	0.048	0.048	0.052
τ_n	ξ for δ_ξ $(Z=1)$													
	0.01	0.02	0.03	0.04	0.07	0.1	0.13	0.16	0.19	0.22	0.25	0.28	0.3	1
0.1	0.153	0.153	0.153	0.145	0.131	0.128	0.112	0.111	0.132	0.142	0.168	0.161	0.174	0.232
0.5	0.084	0.084	0.084	0.083	0.080	0.072	0.067	0.061	0.055	0.052	0.074	0.073	0.066	0.088
1	0.020	0.060	0.077	0.072	0.067	0.061	0.054	0.048	0.051	0.042	0.042	0.042	0.052	0.054
2	0.003	0.031	0.046	0.049	0.063	0.056	0.047	0.037	0.036	0.037	0.026	0.030	0.037	0.037
3	0.000	0.005	0.044	0.048	0.063	0.056	0.047	0.037	0.035	0.037	0.026	0.030	0.037	0.037
4	0.000	0.005	0.035	0.048	0.051	0.056	0.047	0.037	0.035	0.037	0.026	0.030	0.037	0.037
∞	0.000	0.005	0.035	0.048	0.051	0.056	0.047	0.037	0.035	0.037	0.026	0.030	0.037	0.037

Table 15: p-values Obtained from the Proposed Test for Each Measure ν using Application-based Ξ

ξ for δ_ξ										$\bar{\nu}_\xi$
0.03	0.04	0.07	0.1	0.13	0.16	0.19	0.22	0.3	1	
0.957	0.939	0.958	0.975	0.975	0.975	0.975	0.975	0.975	0.975	0.981

References

Aliprantis, C. D. and Border, K. (2006). *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer Science & Business Media.

Bogachev, V. I. (2007). *Measure Theory*, volume 2. Springer Science & Business Media.

Bugni, F. A., Canay, I. A., and Shi, X. (2017). Inference for subvectors and other functions of partially identified parameters in moment inequality models. *Quantitative Economics*, 8(1):1–38.

Card, D. (1993). Using geographic variation in college proximity to estimate the return to schooling. Working Paper 4483, National Bureau of Economic Research.

Davydov, Y. A., Lifshits, M. A., and Smorodina, N. V. (1998). *Local Properties of Distributions of Stochastic Functionals*, volume 173. American Mathematical Society.

Fang, Z. and Santos, A. (2018). Inference on directionally differentiable functions. *The Review of Economic Studies*, 86(1):377–412.

Folland, G. B. (1999). *Real Analysis: Modern Techniques and Their Applications*. John Wiley & Sons.

Giacomini, R., Politis, D. N., and White, H. (2013). A warp-speed method for conducting Monte Carlo experiments involving bootstrap estimators. *Econometric Theory*, 29(3):567–589.

Kitagawa, T. (2015). A test for instrument validity. *Econometrica*, 83(5):2043–2063.

Shapiro, A. (1990). On concepts of directional differentiability. *Journal of Optimization Theory and Applications*, 66(3):477–487.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer.