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

E2023005 2023-2-20

Pairwise Valid Instruments*

Zhenting Sun†
National School of Development
Peking University
zhentingsun@nsd.pku.edu.cn

Kaspar Wüthrich‡
Department of Economics
University of California San Diego
kwuthrich@ucsd.edu

July 19, 2022

Abstract

Finding valid instruments is difficult. We propose Validity Set Instrumental Variable (VSIV) estimation, a method for estimating treatment effects when the instruments are partially invalid. VSIV estimation exploits testable implications of instrument validity to remove invalid variation in the instruments. We show that the proposed VSIV estimators are asymptotically normal under weak conditions and always remove or reduce the asymptotic bias relative to standard IV estimators. We apply VSIV estimation to estimate the returns to schooling using the quarter of birth instrument.

Keywords: Invalid instruments, local average treatment effects, treatment effects, identification, instrumental variable estimation, bias reduction

*We are grateful to Martin Huber, Toru Kitagawa, Julian Martinez-Iriarte, Ismael Mourifié, Xiaoxia Shi, and all seminar participants for their insightful suggestions and comments. The usual disclaimer applies.
†Research supported by the National Natural Science Foundation of China (72103004)
‡Wüthrich is also affiliated with CESifo and the ifo Institute.

1 Introduction

Instrumental variable (IV) methods based on the local average treatment effect (LATE) framework (Imbens and Angrist, 1994; Angrist and Imbens, 1995; Angrist et al., 1996) rely on three assumptions:[1] (i) *exclusion* (the instrument does not have a direct effect on the outcome), (ii) *random assignment* (the instrument is independent of potential outcomes and treatments), and (iii) *monotonicity* (the instrument has a monotonic impact on treatment take-up).[2] In many applications, some of these assumptions are likely to be violated or at least questionable. This has motivated the derivation of testable restrictions and tests for IV validity in various settings (e.g., Balke and Pearl, 1997; Imbens and Rubin, 1997; Heckman and Vytlacil, 2005; Huber and Mellace, 2015; Kitagawa, 2015; Mourifié and Wan, 2017; Kédagni and Mourifié, 2020; Carr and Kitagawa, 2021; Sun, 2021; Farbmacher et al., 2022).[3] The main contribution of this paper is to propose a method for exploiting the information available in the testable restrictions of IV validity to remove or reduce the bias in IV estimation.

We consider a setting where the available instruments are partially invalid. For example, there might be a multivalued instrument for which only some pairs of instrument values satisfy the IV assumptions. In Section 6, we revisit the classical quarter of birth (QOB) instrument of Angrist and Krueger (1991). One potential concern with this instrument is that the seasonality in birth patterns renders some QOBs invalid (e.g., Bound et al., 1995; Buckles and Hungerman, 2013), which motivates some studies to only consider a subset of QOBs as instruments (e.g., Dahl et al., 2017). Our empirical results show that the QOB instrument is indeed partially invalid. Another leading example of partially invalid instruments is when there are several instruments, but only a subset of them are valid.

Our method, which we refer to as *Validity Set IV (VSIV) estimation*, has two steps. First, we use testable implications of IV validity to remove invalid variation in the instruments. Second, we conduct an IV estimation using the remaining variation in the instruments. We establish the asymptotic normality of the proposed VSIV estimators and show that they always remove or reduce the bias relative to traditional IV estimators. Thus, VSIV estimation constitutes a data-driven approach for removing or reducing the bias in IV estimation as much as possible, given all the information about IV validity in the data.

[1] See, for example, Imbens (2014); Melly and Wüthrich (2017); Huber and Wüthrich (2018) for recent reviews, and Angrist and Pischke (2008, 2014); Imbens and Rubin (2015) for textbook treatments.

[2] Some papers also include the instrument first-stage assumption as part of the LATE assumptions. We will maintain suitable first-stage assumptions.

[3] There is a related literature on inference with invalid instruments in linear IV models (e.g., Conley et al., 2012; Armstrong and Kolesár, 2021; Goh and Yu, 2022).

The use of the testable implications of IV validity in VSIV estimation is more constructive than the standard practice where researchers first test for IV validity, discard the instruments if they reject IV validity, and proceed with standard IV analyses if they do not reject IV validity. VSIV estimation uses the testable implications to remove invalid information in the instruments. Consequently, it can be used to estimate causal effects in settings where the instruments are partially invalid so that existing tests reject the null of IV validity. VSIV estimation salvages falsified instruments by exploiting the variation in the instruments not refuted by the data and thereby contributes to the literature on salvaging falsified models (e.g., Kédagni et al., 2020; Masten and Poirier, 2021).

Our goal is to estimate the causal effect of an endogenous treatment D on an outcome of interest Y, using a potentially vector-valued discrete instrument Z. In the ideal case, Z is fully valid, i.e., the LATE assumptions hold for all instrument values (the instrument is valid for the whole population). However, full IV validity is questionable in many applications, especially when there are many instruments or instrument values. To this end, we introduce the notion of *pairwise valid instruments*.[4] Pairwise valid instruments are only valid for a subset of all pairs of instrument values, which we refer to as the *validity pair set*. Intuitively, the instruments are valid for some subpopulations but invalid for the others. For example, as discussed above, not all QOBs might be valid instruments due to the seasonality in birth patterns.

In the first step of VSIV estimation, we identify and estimate the largest validity pair set, $\mathscr{Z}_{\bar{M}}$, using the testable restrictions for IV validity in Kitagawa (2015), Mourifié and Wan (2017), Kédagni and Mourifié (2020), and Sun (2021). In the second step of VSIV estimation, we estimate LATEs for all pairs of instrument values in the estimated validity set, $\widehat{\mathscr{Z}}_0$.

We study the theoretical properties of VSIV estimation under two scenarios. If the estimated validity pair set, $\widehat{\mathscr{Z}}_0$, is consistent for the largest validity pair set $\mathscr{Z}_{\bar{M}}$ in the sense that $\mathbb{P}(\widehat{\mathscr{Z}}_0 = \mathscr{Z}_{\bar{M}}) \to 1$, VSIV estimation is asymptotically unbiased and normal under standard conditions. Since the estimator of the validity pair set, $\widehat{\mathscr{Z}}_0$, is typically constructed based on necessary (but not necessarily sufficient) conditions for IV validity, it could converge to a *pseudo-validity pair set* $\mathscr{Z}_0$ that is larger than $\mathscr{Z}_{\bar{M}}$, i.e., $\mathbb{P}(\widehat{\mathscr{Z}}_0 = \mathscr{Z}_0) \to 1$.[5] We prove that VSIV estimation always leads to a smaller asymptotic bias than standard IV methods. Taken together, our theoretical results show that, irrespective of whether

[4] Pairwise validity can be viewed as a generalization of the partial monotonicity assumption of Mogstad et al. (2021). See Remark 2.2 for a discussion.

[5] Kitagawa (2015, Proposition 1.1) shows that there exist no sufficient conditions for IV validity, even in the simplest case when D and Z are both binary.

the largest validity pair set can be estimated consistently or not, VSIV estimation leads to asymptotically normal IV estimators with reduced bias.

VSIV estimation can be applied in many different settings. In the main text, we focus on the leading case of a binary treatment. In the Appendix, we extend our results to multivalued ordered treatments and also consider unordered treatments (Heckman and Pinto, 2018). Moreover, VSIV estimation is generic—it can be used in conjunction with any set of testable restrictions. For example, if additional testable restrictions beyond those in Kitagawa (2015), Mourifié and Wan (2017), Kédagni and Mourifié (2020), and Sun (2021) become available, they can be used to refine the estimator of the validity pair set $\widehat{\mathscr{Z}_0}$ and further reduce the bias of VSIV estimation.

Notation. We introduce some standard notation (e.g., Sun, 2021). All random elements are defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$. For all $m \in \mathbb{N}$, $\mathcal{B}_{\mathbb{R}^m}$ is the Borel σ-algebra on $\mathbb{R}^m$. We denote by $\mathcal{P}$ the set of probability measures on $(\mathbb{R}^3, \mathcal{B}_{\mathbb{R}^3})$. The symbol $\rightsquigarrow$ denotes weak convergence in a metric space in the Hoffmann–Jørgensen sense. For every subset $B \subset \mathbb{D}$, let 1_B denote the indicator function for B. Finally, we adopt the convention (e.g., Folland, 1999, p. 45), that

$$0 \cdot \infty = 0. \tag{1.1}$$

2 Identification with Pairwise Valid Instruments

Consider a setting with an outcome variable $Y \in \mathbb{R}$, a treatment $D \in \mathcal{D}$, and an instrument (vector) $Z \in \mathcal{Z}$. In the main text, we focus on the leading case where the treatment is binary, $D \in \mathcal{D} = \{0, 1\}$. See the Appendix for extensions to multivalued ordered and unordered treatments. The instrument is discrete, $Z \in \mathcal{Z} = \{z_1, \ldots, z_K\}$, and can be ordered or unordered. Let $Y_{dz} \in \mathbb{R}$ for $(d, z) \in \mathcal{D} \times \mathcal{Z}$ denote the potential outcomes and let D_z for $z \in \mathcal{Z}$ denote the potential treatments. The following assumption generalizes the standard LATE assumptions with binary instruments to multivalued instruments.

Assumption 2.1 *LATE assumptions with binary treatments:*

(i) Exclusion: For each $d \in \{0, 1\}$, $Y_{dz_1} = Y_{dz_2} = \cdots = Y_{dz_K}$ *almost surely (a.s.).*

(ii) Random Assignment: Z *is jointly independent of* $(Y_{0z_1}, \ldots, Y_{0z_K}, Y_{1z_1}, \ldots, Y_{1z_K})$ *and* $(D_{z_1}, \ldots, D_{z_K})$.

(iii) Monotonicity: For all $k = 1, \ldots, K-1$, $D_{z_{k+1}} \geq D_{z_k}$ *a.s.*

Assumption 2.1 is similar to the LATE assumptions in, for example, Imbens and Angrist (1994), Angrist and Imbens (1995), Frölich (2007), Kitagawa (2015), and Sun (2021). It imposes exclusion, random assignment, and monotonicity with respect to all possible values of the instrument $z \in \mathcal{Z}$, which can be restrictive in applications. Therefore, we introduce the notion of *pairwise instrument validity*, which weakens the conditions in Assumption 2.1. Define the set of all possible pairs of values of Z as

$$\mathscr{Z} = \{(z_1, z_2), \ldots, (z_1, z_K), (z_2, z_3), \ldots, (z_2, z_K), \ldots, (z_{K-1}, z_K), (z_2, z_1), \ldots, (z_K, z_{K-1})\}.$$

The number of the elements in $\mathscr{Z}$ is $K \cdot (K-1)$. We use $\mathscr{Z}_{(k,k')}$ to denote a pair $(z_k, z_{k'}) \in \mathscr{Z}$.

Definition 2.1 *The instrument* Z *is* ***pairwise valid*** *for the treatment* $D \in \{0,1\}$ *if there is a set* $\mathscr{Z}_M = \{(z_{k_1}, z_{k'_1}), \ldots, (z_{k_M}, z_{k'_M})\} \subset \mathscr{Z}$ *such that the following conditions hold for every* $(z, z') \in \mathscr{Z}_M$*:*

(i) Exclusion: For each $d \in \{0,1\}$, $Y_{dz} = Y_{dz'}$ *a.s.*

(ii) Random Assignment: Z *is jointly independently of* $(Y_{0z}, Y_{0z'}, Y_{1z}, Y_{1z'}, D_z, D_{z'})$.[6]

(iii) Monotonicity: $D_{z'} \geq D_z$ *a.s.*

The set $\mathscr{Z}_M$ *is called a* ***validity pair set*** *of* Z*. The union of all validity pair sets is the largest validity pair set, denoted by* $\mathscr{Z}_{\bar{M}}$*.*

To illustrate Definition 2.1, consider a simple example where $Z \in \mathcal{Z} = \{z_1, z_2, z_3\}$. If Z is fully valid as in Assumption 2.1 such that $D_{z_3} \geq D_{z_2} \geq D_{z_1}$ a.s., $\mathbb{P}(D_{z_2} > D_{z_1}) > 0$, and $\mathbb{P}(D_{z_3} > D_{z_2}) > 0$, then we have $\mathscr{Z}_{\bar{M}} = \{(z_1, z_2), (z_1, z_3), (z_2, z_3)\}$. The blue solid lines in Figure 2.1(a) indicate that two instrument values, $\{z_k, z_{k'}\}$, form a validity pair: Either $(z_k, z_{k'})$ or $(z_{k'}, z_k)$ satisfies the conditions in Definition 2.1. The full validity Assumption 2.1 requires that every pair of instrument values forms a validity pair. Definition 2.1 relaxes Assumption 2.1 as it does not require every pair to form a validity pair. For example, it could be that only (z_1, z_3) satisfies the conditions in Definition 2.1. The orange dashed lines in Figure 2.1(b) indicate that $\{z_1, z_2\}$ and $\{z_2, z_3\}$ do not form validity pairs. In this case, the instrument Z is pairwise but not fully valid.

[6] This condition can be further weakened: The conditional distribution of $(Y_{0z}, Y_{0z'}, Y_{1z}, Y_{1z'}, D_z, D_{z'})$ given $Z = z$ or $Z = z'$ is the same as the unconditional distribution.

Figure 2.1: Full IV Validity vs. Pairwise IV Validity

(a) Fully Valid Instrument Z (b) Pairwise Valid Instrument Z

Remark 2.1 (Weakening Definition 2.1 with Multiple Instruments) *In Appendix B.2, we introduce a weaker notion of pairwise validity (Definition 2.1) for settings where Z contains multiple instruments: $Z = (Z_1, \ldots, Z_L)^T$, where Z_l is a scalar instrument for $l \in \{1, \ldots, L\}$.*

The following lemma shows that under pairwise IV validity, particular treatment effects can be identified.

Lemma 2.1 *Suppose that the instrument Z is pairwise valid according to Definition 2.1 with a known validity pair set $\mathscr{Z}_M = \{(z_{k_1}, z_{k'_1}), \ldots, (z_{k_M}, z_{k'_M})\}$.*[7] *Then we can define $Y_d(z_{k_m}, z_{k'_m}) = Y_{dz_{k_m}} = Y_{dz_{k'_m}}$ a.s. for each $d \in \{0, 1\}$ and every $(z_{k_m}, z_{k'_m}) \in \mathscr{Z}_M$, and the following quantity can be identified for every $(z_{k_m}, z_{k'_m}) \in \mathscr{Z}_M$:*

$$\begin{aligned} \beta_{k'_m,k_m} &\equiv E\left[Y_1(z_{k_m}, z_{k'_m}) - Y_0(z_{k_m}, z_{k'_m}) \middle| D_{z_{k'_m}} > D_{z_{k_m}}\right] \\ &= \frac{E\left[Y|Z = z_{k'_m}\right] - E\left[Y|Z = z_{k_m}\right]}{E\left[D|Z = z_{k'_m}\right] - E\left[D|Z = z_{k_m}\right]}. \end{aligned} \tag{2.1}$$

Lemma 2.1 is a direct extension of Theorem 1 of Imbens and Angrist (1994) for the case where Z is pairwise valid. We follow Imbens and Angrist (1994) and refer to $\beta_{k'_m,k_m}$ as a LATE. Lemma 2.1 shows that if a validity pair set $\mathscr{Z}_M$ is known, we can identify every $\beta_{k'_m,k_m}$ with $(z_{k_m}, z_{k'_m}) \in \mathscr{Z}_M$. In practice, however, $\mathscr{Z}_M$ is usually unknown. In this paper, we show how to identify and estimate the largest validity pair set $\mathscr{Z}_{\bar{M}}$ based on testable restrictions for IV validity, and how to use the estimated validity pair set to reduce the bias in IV estimation. Note that if $(z_{k_m}, z_{k'_m}) \in \mathscr{Z}_M$ with $D_{z_{k_m}} = D_{z_{k'_m}}$ a.s., then $\beta_{k'_m,k_m} = 0$ by (1.1). Moreover, if $(z_{k_m}, z_{k'_m}) \in \mathscr{Z}_M$ and $(z_{k'_m}, z_{k_m}) \in \mathscr{Z}_M$, then by Definition 2.1 $D_{z_{k_m}} = D_{z_{k'_m}}$ a.s.

We focus on all the LATEs $\beta_{k'_m,k_m}$ as our objects of interest. Traditional IV estimators yield weighted averages of LATEs (e.g., Imbens and Angrist, 1994) and, thus, are strictly

[7] Note that we do not need to impose a first-stage assumption here due to the convention (1.1).

less informative. Moreover, we can always compute linear IV estimands based on the LATEs.

Remark 2.2 (Relationship between Pairwise Validity and Partial Monotonicity) *The partial monotonicity condition proposed by Mogstad et al. (2021) is a special case of Condition (iii) in Definition 2.1; see also Goff (2020) for related assumptions. For example, suppose that* $Z = (Z_1, Z_2) \in \mathbb{R}^2$ *and each element of* Z *is binary so that* $\mathcal{Z} = \{(0,0),(0,1),(1,0),(1,1)\}$. *Suppose that Assumption PM of Mogstad et al. (2021) holds with* $D_{(0,0)} \geq D_{(0,1)}$, $D_{(0,0)} \geq D_{(1,0)}$, $D_{(1,1)} \geq D_{(0,1)}$, *and* $D_{(1,1)} \geq D_{(1,0)}$ *a.s. (the sex composition instrument in Angrist and Evans (1998) discussed in Mogstad et al. (2021)), and that Conditions (i) and (ii) of Definition 2.1 hold. Then a validity pair set is*

$$\{((0,1),(0,0)),((1,0),(0,0)),((0,1),(1,1)),((1,0),(1,1))\}.$$

3 Validity Set IV Estimation

The largest validity pair set $\mathscr{Z}_{\bar{M}}$ is typically unknown in applications. In this paper, we propose a procedure for estimating $\mathscr{Z}_{\bar{M}}$. That is, we seek to identify and exclude $(z_k, z_{k'}) \notin \mathscr{Z}_{\bar{M}}$ from $\mathscr{Z}$, since if $(z_k, z_{k'}) \notin \mathscr{Z}_{\bar{M}}$, then $\beta_{k',k}$ in (2.1) may not be well defined or identified. Suppose that there is a set $\mathscr{Z}_0 \subset \mathscr{Z}$ that satisfies the testable implications in Kitagawa (2015), Mourifié and Wan (2017), Kédagni and Mourifié (2020), and Sun (2021), which we will discuss in detail in Section 4. Then we construct an estimator $\widehat{\mathscr{Z}_0}$ for $\mathscr{Z}_0$. We refer to the IV estimators based on $(z_k, z_{k'}) \in \widehat{\mathscr{Z}_0}$ as *VSIV estimators*. In the following, we assume that a suitable estimator $\widehat{\mathscr{Z}_0}$ is available. We discuss the construction of this estimator in Section 4.

If $\widehat{\mathscr{Z}_0}$ is consistent for the largest validity pair set $\mathscr{Z}_{\bar{M}}$ in the sense that $\mathbb{P}(\widehat{\mathscr{Z}_0} = \mathscr{Z}_{\bar{M}}) \to 1$, the proposed VSIV estimators are asymptotically unbiased and normal under standard weak regularity conditions. We consider this case in Section 3.1. Since $\mathscr{Z}_0$ is constructed based on the necessary (but not necessarily sufficient) conditions for the pairwise IV validity, $\mathscr{Z}_0$ could be larger than $\mathscr{Z}_{\bar{M}}$. (There exist no sufficient conditions for IV validity in general (Kitagawa, 2015).) In Section 3.2, we show that even if $\mathscr{Z}_0$ is larger than $\mathscr{Z}_{\bar{M}}$, VSIV estimators always yield bias reductions relative to standard IV estimators.

3.1 VSIV Estimation under Consistent Estimation of Validity Pair Set

Suppose that the estimator, $\widehat{\mathscr{Z}_0}$, is consistent for the largest validity pair set $\mathscr{Z}_{\bar{M}}$, in the sense that $\mathbb{P}(\widehat{\mathscr{Z}_0} = \mathscr{Z}_{\bar{M}}) \to 1$, and we use $\widehat{\mathscr{Z}_0}$ to construct VSIV estimators for the LATEs. To construct the VSIV estimators and establish their asymptotic properties, we impose the following standard regularity conditions. Let g be a prespecified function that maps the value of Z to $\mathbb{R}$. For example, we can simply set $g(z) = z$ for all z if Z is a scalar instrument.[8]

Assumption 3.1 *$\{(Y_i, D_i, Z_i)\}_{i=1}^n$ is an i.i.d. sample from a population such that all relevant moments exist.*

Assumption 3.2 *For every $\mathcal{Z}_{(k,k')} \in \mathscr{Z}_{\bar{M}}$,*

$$E[g(Z_i)D_i|Z_i \in \mathcal{Z}_{(k,k')}] - E[D_i|Z_i \in \mathcal{Z}_{(k,k')}] \cdot E[g(Z_i)|Z_i \in \mathcal{Z}_{(k,k')}] \neq 0. \tag{3.1}$$

Assumption 3.1 assumes an i.i.d. data set and requires the existence of the relevant moments. Assumption 3.2 imposes a first-stage condition for every $\mathcal{Z}_{(k,k')} \in \mathscr{Z}_{\bar{M}}$. Note that (3.1) may not hold for $\mathcal{Z}_{(k,k')} \notin \mathscr{Z}_{\bar{M}}$. This creates additional technical difficulties when establishing the asymptotic normality of the VSIV estimators, which we discuss below. Assumption 3.2 also implies that if $\mathcal{Z}_{(k,k')} \in \mathscr{Z}_{\bar{M}}$, then $\mathcal{Z}_{(k',k)} \notin \mathscr{Z}_{\bar{M}}$. Otherwise, by Definition 2.1, $D_{z_k} = D_{z_{k'}}$ and (3.1) does not hold. For every random variable ξ_i and every $\mathcal{A} \in \mathscr{Z}$, we define

$$\mathcal{E}_n(\xi_i, \mathcal{A}) = \frac{\frac{1}{n}\sum_{i=1}^n \xi_i 1\{Z_i \in \mathcal{A}\}}{\frac{1}{n}\sum_{i=1}^n 1\{Z_i \in \mathcal{A}\}} \text{ and } \mathcal{E}(\xi_i, \mathcal{A}) = \frac{E[\xi_i 1\{Z_i \in \mathcal{A}\}]}{E[1\{Z_i \in \mathcal{A}\}]}.$$

For every $\mathcal{Z}_{(k,k')} \in \mathscr{Z}$, we run the IV regression

$$Y_i 1\{Z_i \in \mathcal{Z}_{(k,k')}\} = \gamma^0_{(k,k')} 1\{Z_i \in \mathcal{Z}_{(k,k')}\} + \gamma^1_{(k,k')} D_i 1\{Z_i \in \mathcal{Z}_{(k,k')}\} + \epsilon_i 1\{Z_i \in \mathcal{Z}_{(k,k')}\}, \tag{3.2}$$

using $g(Z_i)1\{Z_i \in \mathcal{Z}_{(k,k')}\}$ as the instrument for $D_i 1\{Z_i \in \mathcal{Z}_{(k,k')}\}$. Given the estimated validity set $\widehat{\mathscr{Z}_0}$, we set the VSIV estimator for each $\mathcal{Z}_{(k,k')}$ as

$$\widehat{\beta}^1_{(k,k')} = 1\left\{\mathcal{Z}_{(k,k')} \in \widehat{\mathscr{Z}_0}\right\} \cdot \frac{\mathcal{E}_n\left(g(Z_i) Y_i, \mathcal{Z}_{(k,k')}\right) - \mathcal{E}_n\left(g(Z_i), \mathcal{Z}_{(k,k')}\right)\mathcal{E}_n\left(Y_i, \mathcal{Z}_{(k,k')}\right)}{\mathcal{E}_n\left(g(Z_i) D_i, \mathcal{Z}_{(k,k')}\right) - \mathcal{E}_n\left(g(Z_i), \mathcal{Z}_{(k,k')}\right)\mathcal{E}_n\left(D_i, \mathcal{Z}_{(k,k')}\right)}, \tag{3.3}$$

[8] The choice of g will affect the efficiency of the VSIV estimators. We leave the formal analysis of the optimal choice of g for future study.

which is the IV estimator of $\gamma^1_{(k,k')}$ in (3.2) multiplied by $1\{\mathcal{Z}_{(k,k')} \in \widehat{\mathscr{Z}_0}\}$.

Remark 3.1 *For every $\mathcal{Z}_{(k,k')} \in \widehat{\mathscr{Z}_0}$, the estimation in (3.3) is equivalent to the canonical IV estimation in the subsample of $\{(Y_i, D_i, Z_i)\}_{i=1}^n$ with $Z_i \in \mathcal{Z}_{(k,k')}$.*

Define the vector of VSIV estimators as

$$\widehat{\beta}_1 = \left(\widehat{\beta}^1_{(1,2)}, \ldots, \widehat{\beta}^1_{(1,K)}, \ldots, \widehat{\beta}^1_{(K,1)}, \ldots, \widehat{\beta}^1_{(K,K-1)}\right)^T.$$

We also define

$$\beta^1_{(k,k')} = 1\left\{\mathcal{Z}_{(k,k')} \in \mathscr{Z}_{\bar{M}}\right\} \cdot \frac{\mathcal{E}\left(g\left(Z_i\right) Y_i, \mathcal{Z}_{(k,k')}\right) - \mathcal{E}\left(g\left(Z_i\right), \mathcal{Z}_{(k,k')}\right) \mathcal{E}\left(Y_i, \mathcal{Z}_{(k,k')}\right)}{\mathcal{E}\left(g\left(Z_i\right) D_i, \mathcal{Z}_{(k,k')}\right) - \mathcal{E}\left(g\left(Z_i\right), \mathcal{Z}_{(k,k')}\right) \mathcal{E}\left(D_i, \mathcal{Z}_{(k,k')}\right)} \tag{3.4}$$

and

$$\beta_1 = \left(\beta^1_{(1,2)}, \ldots, \beta^1_{(1,K)}, \ldots, \beta^1_{(K,1)}, \ldots, \beta^1_{(K,K-1)}\right)^T. \tag{3.5}$$

Remark 3.2 *If $\mathcal{Z}_{(k,k')} \notin \mathscr{Z}_{\bar{M}}$, the parameter $\beta_{k',k}$ in (2.1) may not be well defined or identified, and we set $\beta^1_{(k,k')} = 0$ by (3.4) and (1.1). Similarly, if $\mathcal{Z}_{(k,k')} \notin \widehat{\mathscr{Z}_0}$, $\widehat{\beta}^1_{(k,k')} = 0$ by (3.3) and (1.1).*

The next theorem establishes the asymptotic distribution of the vector of VSIV estimator $\widehat{\beta}_1$, obtained based on the estimator of the instrument validity pair set $\widehat{\mathscr{Z}_0}$.

Theorem 3.1 *Suppose that the instrument Z is pairwise valid for the treatment D according to Definition 2.1 with the largest validity pair set $\mathscr{Z}_{\bar{M}} = \{(z_{k_1}, z_{k'_1}), \ldots, (z_{k_{\bar{M}}}, z_{k'_{\bar{M}}})\}$, and that the estimator $\widehat{\mathscr{Z}_0}$ satisfies $\mathbb{P}(\widehat{\mathscr{Z}_0} = \mathscr{Z}_{\bar{M}}) \to 1$. Under Assumptions 3.1 and 3.2,*

$$\sqrt{n}(\widehat{\beta}_1 - \beta_1) \xrightarrow{d} N\left(0, \Sigma\right), \tag{3.6}$$

where Σ is defined in (B.5) in the Appendix. In addition, $\beta^1_{(k,k')} = \beta_{k',k}$ as defined in (2.1) for every $(z_k, z_{k'}) \in \mathscr{Z}_{\bar{M}}$.

Theorem 3.1 establishes the joint asymptotic normality of the VSIV estimator of the LATEs. Establishing the asymptotic distribution in (3.6) requires a careful treatment of the case where the first-stage Assumption 3.2 does not hold for some pairs of instrument values $\mathcal{Z}_{(k,k')}$ that are not in the largest validity pair set $\mathscr{Z}_{\bar{M}}$, that is, $\mathcal{Z}_{(k,k')} \notin \mathscr{Z}_{\bar{M}}$. Specifically,

we show that in this case, $\mathbb{P}(\widehat{\mathscr{Z}_0} = \mathscr{Z}_{\bar{M}}) \to 1$ implies that, if $\mathcal{Z}_{(k,k')} \notin \mathscr{Z}_{\bar{M}}$, then for every $\rho > 0$, $n^\rho 1\{\mathcal{Z}_{(k,k')} \in \widehat{\mathscr{Z}_0}\} = o_p(1)$. This guarantees the convergence in (3.6) even when Assumption 3.2 does not hold for $\mathcal{Z}_{(k,k')}$. The asymptotic covariance matrix Σ defined in the Appendix can be consistently estimated under standard conditions. Importantly, the estimation of the instrument validity pair set does not affect the asymptotic covariance matrix such that standard inference methods can be applied.

The LATE $\beta_{k',k}$ may not be identified if $\mathcal{Z}_{(k,k')} \notin \mathscr{Z}_{\bar{M}}$. Let $\beta_{1S} = (\beta^1_{(\kappa_1,\kappa'_1)}, \ldots, \beta^1_{(\kappa_S,\kappa'_S)})^T$ for some $S > 0$. In our context, it is interesting to test hypotheses about $\beta^1_{(k,k')}$ with $\mathcal{Z}_{(k,k')} \in \mathscr{Z}_{\bar{M}}$ ($\beta^1_{(k,k')}$ is equal to the LATE $\beta_{k',k}$ by Theorem 3.1):

$$\mathrm{H}_0 : \mathcal{Z}_{(\kappa_1,\kappa'_1)} \in \mathscr{Z}_{\bar{M}}, \ldots, \mathcal{Z}_{(\kappa_S,\kappa'_S)} \in \mathscr{Z}_{\bar{M}}, \ R(\beta_{1S}) = 0, \tag{3.7}$$

where R is a (possibly nonlinear) smooth r-dimensional function. Let $R'(\beta_S)$ be the $r \times S$ matrix of the continuous first derivative functions at an arbitrary value β_S, that is, $R'(\beta_S) = \partial R(\beta_S)/\partial \beta_S^T$. Let $\mathcal{I}_S$ be a $S \times (K-1)K$ matrix such that for every $\beta = (\beta_{(1,2)}, \ldots, \beta_{(1,K)}, \ldots, \beta_{(K,1)}, \ldots, \beta_{(K,K-1)})^T$,

$$\mathcal{I}_S \beta = (\beta_{(\kappa_1,\kappa'_1)}, \ldots, \beta_{(\kappa_S,\kappa'_S)})^T.$$

Theorem 3.1 implies that

$$\sqrt{n}(\widehat{\beta}_{1S} - \beta_{1S}) = \sqrt{n}\mathcal{I}_S(\widehat{\beta}_1 - \beta_1) \xrightarrow{d} N(0, \Sigma_S),$$

where $\Sigma_S = \mathcal{I}_S \Sigma \mathcal{I}_S^T$ so that, by the delta method, we obtain

$$\sqrt{n}\left\{R(\widehat{\beta}_{1S}) - R(\beta_{1S})\right\} \xrightarrow{d} N\left(0, R'(\beta_{1S})\Sigma_S R'(\beta_{1S})^T\right).$$

We construct the test statistics as

$$TS_{1n} = \prod_{s=1}^{S} 1\left\{\mathcal{Z}_{(\kappa_s,\kappa'_s)} \in \widehat{\mathscr{Z}_0}\right\}$$

and

$$TS_{2n} = \sqrt{n}R(\widehat{\beta}_{1S})^T \left\{R'(\widehat{\beta}_{1S})\mathcal{I}_S \widehat{\Sigma} \mathcal{I}_S^T R'(\widehat{\beta}_{1S})^T\right\}^{-1} \sqrt{n}R(\widehat{\beta}_{1S}),$$

where $\widehat{\Sigma}$ is a consistent estimator of Σ, which can be constructed based on the formula in (B.5). Suppose that Assumptions 3.1 and 3.2 hold and $\mathbb{P}(\widehat{\mathscr{Z}_0} = \mathscr{Z}_{\bar{M}}) \to 1$. If H_0 is true and $R'(\beta_{1S})$ is of full row rank, then it follows from standard arguments that $TS_{2n} \xrightarrow{d} \chi^2_r$ for

some chi-square distribution χ^2_r with the degrees of freedom r. The decision rule of the test is to reject H_0 if $TS_{1n} = 0$ or $TS_{2n} > c_r(\alpha)$, where $c_r(\alpha)$ is such that $\mathbb{P}(\chi^2_r > c_r(\alpha)) = \alpha$ for some predetermined $\alpha \in (0, 1)$. The following proposition establishes the formal properties of the proposed test.

Proposition 3.1 *Suppose that Assumptions 3.1 and 3.2 hold and* $\mathbb{P}(\widehat{\mathscr{Z}_0} = \mathscr{Z}_{\bar{M}}) \to 1$.

(i) If H_0 *is true,* $\mathbb{P}\left(\{TS_{1n} = 0\} \cup \{TS_{2n} > c_r(\alpha)\}\right) \to \alpha$.

(ii) If H_0 *is false,* $\mathbb{P}\left(\{TS_{1n} = 0\} \cup \{TS_{2n} > c_r(\alpha)\}\right) \to 1$.

3.2 Bias Reduction using VSIV Estimation

In Section 3.1, we show that if the estimator of the validity set is consistent, $\mathbb{P}(\widehat{\mathscr{Z}_0} = \mathscr{Z}_{\bar{M}}) \to 1$, VSIV estimators are consistent for LATEs under weak conditions. However, since $\mathscr{Z}_0$ is constructed based on necessary (but not necessarily sufficient) conditions for IV validity, we have $\mathbb{P}(\widehat{\mathscr{Z}_0} = \mathscr{Z}_0) \to 1$ in general, where the *pseudo-validity pair set* $\mathscr{Z}_0$ could be larger than $\mathscr{Z}_{\bar{M}}$. In this case, VSIV may not be asymptotically unbiased. Here we show that even if $\mathscr{Z}_0$ is larger than $\mathscr{Z}_{\bar{M}}$, the VSIV estimators always reduce the bias relative to standard IV estimators.[9] Intuitively, VSIV estimators use the information in the data about IV validity to reduce the asymptotic bias as much as possible.

Since our target parameter is the vector β_1, a natural definition of the estimation bias is $\|\tilde{\beta}_1 - \beta_1\|_2$ for every estimator $\tilde{\beta}_1$.

Definition 3.1 *The estimation bias of an arbitrary estimator* $\tilde{\beta}_1$ *for the true value* β_1 *defined in* (3.5) *is defined as* $\|\tilde{\beta}_1 - \beta_1\|_2$*, where* $\|\cdot\|_2$ *is the* ℓ^2*-norm on Euclidean spaces.*

Consider an arbitrary presumed validity pair set $\mathscr{Z}_P$, which could incorporate prior information. Given $\mathscr{Z}_P$, we define $\widehat{\mathscr{Z}_0'} = \widehat{\mathscr{Z}_0} \cap \mathscr{Z}_P$ and use $\widehat{\mathscr{Z}_0'}$ to construct the VSIV estimators in (3.3).

The next assumption extends Assumption 3.2 to $\mathscr{Z}_0$.

Assumption 3.3 *For every* $\mathcal{Z}_{(k,k')} \in \mathscr{Z}_0$*,*

$$E[g(Z_i)D_i | Z_i \in \mathcal{Z}_{(k,k')}] - E[D_i | Z_i \in \mathcal{Z}_{(k,k')}] \cdot E[g(Z_i) | Z_i \in \mathcal{Z}_{(k,k')}] \neq 0. \tag{3.8}$$

[9] Standard IV estimators are equal to VSIV estimators with some presumed (unverifiable) validity pair set.

The following theorem shows that the VSIV estimators based on $\widehat{\mathscr{Z}_0'}$ always exhibit a smaller asymptotic bias than standard IV estimators based on $\mathscr{Z}_P$.

Theorem 3.2 *Suppose that Assumptions 3.1 and 3.3 hold and that* $\mathbb{P}(\widehat{\mathscr{Z}_0} = \mathscr{Z}_0) \to 1$ *with* $\mathscr{Z}_0 \supset \mathscr{Z}_{\bar{M}}$. *For every presumed validity pair set* $\mathscr{Z}_P$, *the asymptotic estimation bias* $\operatorname{plim}_{n\to\infty}\|\widehat{\beta}_1 - \beta_1\|_2$ *is always reduced by using* $\widehat{\mathscr{Z}_0'}$ *in the estimation* (3.3) *compared to the bias from using* $\mathscr{Z}_P$.

As shown in Proposition 4.1 below, the pseudo-validity pair set $\mathscr{Z}_0$ can always be estimated consistently by $\widehat{\mathscr{Z}_0}$ under mild conditions. Compared to constructing standard IV estimators based on $\mathscr{Z}_P$, Theorem 3.2 shows that the asymptotic estimation bias, $\operatorname{plim}_{n\to\infty}\|\widehat{\beta}_1 - \beta_1\|_2$, can be reduced by using VSIV estimators based on $\widehat{\mathscr{Z}_0'} = \widehat{\mathscr{Z}_0} \cap \mathscr{Z}_P$.

The arguments used for establishing the asymptotic normality of the VSIV estimators in Section 3.1 do not rely on the consistent estimation of $\mathscr{Z}_{\bar{M}}$. Thus, irrespective of whether $\mathscr{Z}_{\bar{M}}$ can be estimated consistently, the VSIV estimators are asymptotically normal, centered at β_1 defined with $\mathscr{Z}_0$ instead of $\mathscr{Z}_{\bar{M}}$. However, note that β_1 can only be interpreted as a vector of LATEs under consistent estimation.

Example 3.1 (Bias Reduction using VSIV Estimation) *Consider a simple example where* $\mathcal{Z} = \{1,2,3,4\}$ *as in our application and suppose that* $\mathscr{Z}_{\bar{M}} = \{(1,2)\}$. *In this case, by* (3.4) *and* (1.1),

$$\beta_1 = \left(\beta^1_{(1,2)}, \ldots, \beta^1_{(1,4)}, \ldots, \beta^1_{(4,1)}, \ldots, \beta^1_{(4,3)}\right)^T = \left(\beta^1_{(1,2)}, 0, \ldots, 0\right)^T.$$

Suppose that, by mistake, we assume Z *is valid according to Assumption 2.1 and use*

$$\mathscr{Z}_P = \{(1,2),(1,3),(1,4),(2,3),(2,4),(3,4)\}$$

as an estimator for $\mathscr{Z}_{\bar{M}}$. *Then by* (3.3) *and* (1.1),

$$\widehat{\beta}_1 = \left(\widehat{\beta}^1_{(1,2)}, \widehat{\beta}^1_{(1,3)}, \widehat{\beta}^1_{(1,4)}, \widehat{\beta}^1_{(2,3)}, \widehat{\beta}^1_{(2,4)}, \widehat{\beta}^1_{(3,4)}, 0, 0, 0, 0, 0, 0\right)^T, \tag{3.9}$$

where $\widehat{\beta}^1_{(1,3)}$, $\widehat{\beta}^1_{(1,4)}$, $\widehat{\beta}^1_{(2,3)}$, $\widehat{\beta}^1_{(2,4)}$, *and* $\widehat{\beta}^1_{(3,4)}$ *may not converge to* 0 *in probability. However, by definition* $\beta^1_{(1,3)} = 0$, $\beta^1_{(1,4)} = 0$, $\beta^1_{(2,3)} = 0$, $\beta^1_{(2,4)} = 0$, *and* $\beta^1_{(3,4)} = 0$. *Thus, the bias* $\|\widehat{\beta}_1 - \beta_1\|_2$ *may not converge to* 0 *in probability. The approach proposed in this paper helps reducing this bias as much as possible. We exploit the information in the data about IV validity to obtain the estimator* $\widehat{\mathscr{Z}_0}$. *Even if* $\widehat{\mathscr{Z}_0}$ *converges to a set larger than* $\mathscr{Z}_{\bar{M}}$ *(because we use the*

necessary but not sufficient conditions for IV validity), VSIV always reduces the bias. Suppose that $\mathscr{Z}_0 = \{(1,2),(3,4)\}$*, which is larger than* $\mathscr{Z}_{\bar{M}}$ *but smaller than* $\mathscr{Z}_P$*. In this case, the VSIV estimator* $\widehat{\beta}_1$ *constructed by using* $\widehat{\mathscr{Z}_0} \cap \mathscr{Z}_P$ *converges in probability to*

$$\beta_1' = \left(\beta^1_{(1,2)}, 0, 0, 0, 0, \beta^{1\prime}_{(3,4)}, 0, 0, 0, 0, 0, 0\right)^T, \tag{3.10}$$

where $\beta^{1\prime}_{(3,4)}$ *is the probability limit of* $\widehat{\beta}^1_{(3,4)}$*. Then, clearly, VSIV reduces the probability limit of the bias* $\|\widehat{\beta}_1 - \beta_1\|_2$*.*

3.3 Partially Valid Instruments and Connection to Existing Results

Suppose we estimate the following canonical IV regression model,

$$Y_i = \alpha_0 + \alpha_1 D_i + \epsilon_i, \tag{3.11}$$

using $g(Z_i)$ as the instrument for D_i. When the instrument Z is fully valid, the traditional IV estimator of α_1 is

$$\widehat{\alpha}_1 = \frac{n\sum_{i=1}^n g\left(Z_i\right) Y_i - \sum_{i=1}^n g\left(Z_i\right) \sum_{i=1}^n Y_i}{n\sum_{i=1}^n g\left(Z_i\right) D_i - \sum_{i=1}^n g\left(Z_i\right) \sum_{i=1}^n D_i}. \tag{3.12}$$

The asymptotic properties of $\widehat{\alpha}_1$ can be found in Imbens and Angrist (1994, p. 471) and Angrist and Imbens (1995, p. 436).

To connect VSIV estimation to canonical IV regression with fully valid instruments, consider the following special case of pairwise IV validity.

Definition 3.2 *Suppose that the instrument* Z *is pairwise valid for the treatment* D *with the largest validity pair set* $\mathscr{Z}_{\bar{M}}$*. If there is a validity pair set*

$$\mathscr{Z}_M = \{(z_{k_1}, z_{k_2}), (z_{k_2}, z_{k_3}), \ldots, (z_{k_{M-1}}, z_{k_M})\}$$

for some $M > 0$*, then the instrument* Z *is called a* ***partially valid instrument*** *for the treatment* D*. The set* $\mathcal{Z}_M = \{z_{k_1}, \ldots, z_{k_M}\}$ *is called a* ***validity value set*** *of* Z*.*

Assumption 3.4 *The validity value set* $\mathcal{Z}_M$ *satisfies that*

$$E[g(Z_i)D_i | Z_i \in \mathcal{Z}_M] - E[D_i | Z_i \in \mathcal{Z}_M] \cdot E[g(Z_i) | Z_i \in \mathcal{Z}_M] \neq 0. \tag{3.13}$$

Suppose that Z is partially valid for the treatment D with a validity value set $\mathcal{Z}_M$, and that there is a consistent estimator $\widehat{\mathcal{Z}_0}$ of $\mathcal{Z}_M$. We then construct a VSIV estimator for α_1 in (3.11) by running the IV estimation for the model

$$Y_i 1\left\{Z_i \in \widehat{\mathcal{Z}_0}\right\} = \gamma_0 1\left\{Z_i \in \widehat{\mathcal{Z}_0}\right\} + \gamma_1 D_i 1\left\{Z_i \in \widehat{\mathcal{Z}_0}\right\} + \epsilon_i 1\left\{Z_i \in \widehat{\mathcal{Z}_0}\right\}, \tag{3.14}$$

using $g(Z_i)1\{Z_i \in \widehat{\mathcal{Z}_0}\}$ as the instrument for $D_i 1\{Z_i \in \widehat{\mathcal{Z}_0}\}$. We obtain the VSIV estimator for α_1 in (3.11) by

$$\widehat{\theta}_1 = \frac{n_z \sum_{i=1}^n g\left(Z_i\right) Y_i 1\left\{Z_i \in \widehat{\mathcal{Z}_0}\right\} - \sum_{i=1}^n g\left(Z_i\right) 1\left\{Z_i \in \widehat{\mathcal{Z}_0}\right\} \sum_{i=1}^n Y_i 1\left\{Z_i \in \widehat{\mathcal{Z}_0}\right\}}{n_z \sum_{i=1}^n g\left(Z_i\right) D_i 1\left\{Z_i \in \widehat{\mathcal{Z}_0}\right\} - \sum_{i=1}^n g\left(Z_i\right) 1\left\{Z_i \in \widehat{\mathcal{Z}_0}\right\} \sum_{i=1}^n D_i 1\left\{Z_i \in \widehat{\mathcal{Z}_0}\right\}}, \tag{3.15}$$

where $n_z = \sum_{i=1}^n 1\{Z_i \in \widehat{\mathcal{Z}_0}\}$. We can see that $\widehat{\theta}_1$ is a generalized version of $\widehat{\alpha}_1$ in (3.12), because when the instrument is fully valid, we can just let $\widehat{\mathcal{Z}_0} = \mathcal{Z}$ and then $\widehat{\theta}_1 = \widehat{\alpha}_1$.

Theorem 3.3 *Suppose that the instrument Z is partially valid for the treatment D according to Definition 3.2 with a validity value set $\mathcal{Z}_M = \{z_{k_1}, \ldots, z_{k_M}\}$, and that the estimator $\widehat{\mathcal{Z}_0}$ for $\mathcal{Z}_M$ satisfies $\mathbb{P}(\widehat{\mathcal{Z}_0} = \mathcal{Z}_M) \to 1$. Under Assumptions 3.1 and 3.4, it follows that $\widehat{\theta}_1 \xrightarrow{p} \theta_1$, where*

$$\theta_1 = \frac{E\left[g\left(Z_i\right) Y_i | Z_i \in \mathcal{Z}_M\right] - E\left[Y_i | Z_i \in \mathcal{Z}_M\right] E\left[g\left(Z_i\right) | Z_i \in \mathcal{Z}_M\right]}{E\left[g\left(Z_i\right) D_i | Z_i \in \mathcal{Z}_M\right] - E\left[D_i | Z_i \in \mathcal{Z}_M\right] E\left[g\left(Z_i\right) | Z_i \in \mathcal{Z}_M\right]}.$$

Also, $\sqrt{n}(\widehat{\theta}_1 - \theta_1) \xrightarrow{d} N\left(0, \Sigma_1\right)$, where Σ_1 is provided in (B.25) in the Appendix. In addition, the quantity θ_1 can be interpreted as the weighted average of $\{\beta_{k_2,k_1}, \ldots, \beta_{k_M,k_{M-1}}\}$ defined as in (2.1). Specifically, $\theta_1 = \sum_{m=1}^{M-1} \mu_m \beta_{k_{m+1},k_m}$ with

$$\mu_m = \frac{\left[p\left(z_{k_{m+1}}\right) - p\left(z_{k_m}\right)\right] \sum_{l=m}^{M-1} \mathbb{P}\left(Z_i = z_{k_{l+1}} | Z_i \in \mathcal{Z}_M\right) \left\{g\left(z_{k_{l+1}}\right) - E\left[g\left(Z_i\right) | Z_i \in \mathcal{Z}_M\right]\right\}}{\sum_{l=1}^{M} \mathbb{P}\left(Z_i = z_{k_l} | Z_i \in \mathcal{Z}_M\right) p\left(z_{k_l}\right) \left\{g\left(z_{k_l}\right) - E\left[g\left(Z_i\right) | Z_i \in \mathcal{Z}_M\right]\right\}},$$

$p\left(z_k\right) = E\left[D_i | Z_i = z_k\right]$, and $\sum_{m=1}^{M-1} \mu_m = 1$.

Theorem 3.3 is an extension of Theorem 2 of Imbens and Angrist (1994) to the case where the instrument is partially but not fully valid. To establish a connection to existing results, Theorem 3.3 assumes consistent estimation of the validity value set, $\mathbb{P}(\widehat{\mathcal{Z}_0} = \mathcal{Z}_M) \to 1$. If $\widehat{\mathcal{Z}_0}$ converges to a larger set than $\mathcal{Z}_M$, the properties of VSIV follow from

the results in Section 3.2 because partially valid instruments are a special case of pairwise valid instruments.

4 Definition and Estimation of $\mathscr{Z}_0$

Here we discuss the definition and the estimation of $\mathscr{Z}_0$ based on the testable implications in Kitagawa (2015), Mourifié and Wan (2017), Kédagni and Mourifié (2020), and Sun (2021) for pairwise IV validity. We show that under weak assumptions, the proposed estimator $\widehat{\mathscr{Z}_0}$ is consistent for the pseudo-validity set $\mathscr{Z}_0$ in the sense that $\mathbb{P}(\widehat{\mathscr{Z}_0} = \mathscr{Z}_0) \to 1$. As a consequence, when $\mathscr{Z}_0 = \mathscr{Z}_{\bar{M}}$, the largest validity pair set can be estimated consistently.

When D is binary, by Lemma B.1, the testable implications in Kédagni and Mourifié (2020) are implied by those in Kitagawa (2015), Mourifié and Wan (2017), and Sun (2021), and we focus on the latter testable implications throughout this section.[10] We are not aware of results on the connection between these two sets of testable implications with multivalued D. Therefore, when D is multivalued, we construct two sets of pairs of instrument values satisfying the testable implications in Kitagawa (2015), Mourifié and Wan (2017), Sun (2021), and those in Kédagni and Mourifié (2020), respectively, and construct $\mathscr{Z}_0$ as the intersection of these two sets (see Appendices B.4 and C.2).

The definition of $\mathscr{Z}_0$ relies on the testable implications proposed in Kitagawa (2015), Mourifié and Wan (2017), and Sun (2021). These testable implications were originally proposed for full IV validity. In the following, we extend them to Definition 2.1. To describe the testable restrictions, we use the notation of Sun (2021). Define conditional probabilities

$$P_z(B, C) = \mathbb{P}(Y \in B, D \in C | Z = z)$$

for all Borel sets $B, C \in \mathcal{B}_{\mathbb{R}}$ and all $z \in \mathcal{Z}$. With the largest validity pair set $\mathscr{Z}_{\bar{M}} = \{(z_{k_1}, z_{k'_1}), \ldots, (z_{k_{\bar{M}}}, z_{k'_{\bar{M}}})\}$, for every $m \in \{1, \ldots, \bar{M}\}$, it follows that

$$P_{z_{k_m}}(B, \{1\}) \leq P_{z_{k'_m}}(B, \{1\}) \text{ and } P_{z_{k_m}}(B, \{0\}) \geq P_{z_{k'_m}}(B, \{0\}) \tag{4.1}$$

for all $B \in \mathcal{B}_{\mathbb{R}}$. By definition, for all $B, C \in \mathcal{B}_{\mathbb{R}}$,

$$\mathbb{P}(Y \in B, D \in C | Z = z) = \frac{\mathbb{P}(Y \in B, D \in C, Z = z)}{\mathbb{P}(Z = z)}.$$

[10] We note that this result is tailored to our focus on LATE-style parameters. For other parameters of interest, it is possible that the testable restrictions in Kédagni and Mourifié (2020) can help obtain sharper identification results. We thank Ismael Mourifié for pointing this out to us.

Define the function spaces

$$\begin{aligned}
\mathcal{G}_P &= \left\{\left(1_{\mathbb{R}\times\mathbb{R}\times\{z_k\}}, 1_{\mathbb{R}\times\mathbb{R}\times\{z_{k'}\}}\right) : k, k' \in \{1,\ldots,K\}, k \neq k'\right\}, \\
\mathcal{H} &= \left\{(-1)^d \cdot 1_{B\times\{d\}\times\mathbb{R}} : B \text{ is a closed interval in } \mathbb{R}, d \in \{0,1\}\right\}, \text{ and} \\
\bar{\mathcal{H}} &= \left\{(-1)^d \cdot 1_{B\times\{d\}\times\mathbb{R}} : B \text{ is a closed, open, or half-closed interval in } \mathbb{R}, d \in \{0,1\}\right\}.
\end{aligned} \tag{4.2}$$

Similarly to Sun (2021), by Lemma B.7 in Kitagawa (2015), we use all closed intervals $B \subset \mathbb{R}$ to construct $\mathcal{H}$ instead of all Borel sets.

Suppose we have access to an i.i.d. sample $\{(Y_i, D_i, Z_i)\}_{i=1}^n$ distributed according to some probability distribution P in $\mathcal{P}$, that is, $P(G) = \mathbb{P}((Y_i, D_i, Z_i) \in G)$ for all $G \in \mathcal{B}_{\mathbb{R}^3}$. For every measurable function v, with some abuse of notation, define

$$P(v) = \int v \, \mathrm{d}P.$$

The closure of $\mathcal{H}$ in $L^2(P)$ is equal to $\bar{\mathcal{H}}$ by Lemma C.1 of Sun (2021). For every $(h, g) \in \bar{\mathcal{H}} \times \mathcal{G}_P$ with $g = (g_1, g_2)$, define

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

and

$$\sigma^2(h, g) = \Lambda(P) \cdot \left\{\frac{P(h^2 \cdot g_2)}{P^2(g_2)} - \frac{P^2(h \cdot g_2)}{P^3(g_2)} + \frac{P(h^2 \cdot g_1)}{P^2(g_1)} - \frac{P^2(h \cdot g_1)}{P^3(g_1)}\right\}, \tag{4.3}$$

where $\Lambda(P) = \prod_{k=1}^K P(1_{\mathbb{R}\times\mathbb{R}\times\{z_k\}})$ and $P^m(g_j) = [P(g_j)]^m$ for $m \in \mathbb{N}$ and $j \in \{1, 2\}$. We denote the sample analog of ϕ as

$$\widehat{\phi}(h, g) = \frac{\widehat{P}(h \cdot g_2)}{\widehat{P}(g_2)} - \frac{\widehat{P}(h \cdot g_1)}{\widehat{P}(g_1)},$$

where $\widehat{P}$ is the empirical probability measure corresponding to P so that for every measurable function v,

$$\widehat{P}(v) = \frac{1}{n}\sum_{i=1}^n v(Y_i, D_i, Z_i). \tag{4.4}$$

For every $(h, g) \in \bar{\mathcal{H}} \times \mathcal{G}_P$ with $g = (g_1, g_2)$, define the sample analog of $\sigma^2(h, g)$ as

$$\widehat{\sigma}^2(h, g) = \frac{T_n}{n} \cdot \left\{ \frac{\widehat{P}(h^2 \cdot g_2)}{\widehat{P}^2(g_2)} - \frac{\widehat{P}^2(h \cdot g_2)}{\widehat{P}^3(g_2)} + \frac{\widehat{P}(h^2 \cdot g_1)}{\widehat{P}^2(g_1)} - \frac{\widehat{P}^2(h \cdot g_1)}{\widehat{P}^3(g_1)} \right\},$$

where $T_n = n \cdot \prod_{k=1}^{K} \widehat{P}(1_{\mathbb{R}\times\mathbb{R}\times\{z_k\}})$. By (1.1), $\widehat{\sigma}^2$ is well defined. By similar proof of Lemma 3.1 in Sun (2021), σ^2 and $\widehat{\sigma}^2$ are uniformly bounded in (h, g). The following lemma reformulates the testable restrictions in (4.1) in terms of ϕ. Below, we use this reformulation to define $\mathscr{Z}_0$ and the corresponding estimator $\widehat{\mathscr{Z}_0}$.

Lemma 4.1 *Suppose that the instrument Z is pairwise valid for the treatment D with the largest validity pair set $\mathscr{Z}_{\bar{M}} = \{(z_{k_1}, z_{k'_1}), \ldots, (z_{k_{\bar{M}}}, z_{k'_{\bar{M}}})\}$. For every $m \in \{1, \ldots, \bar{M}\}$, we have that $\sup_{h\in\mathcal{H}} \phi(h, g) = 0$ with $g = (1_{\mathbb{R}\times\mathbb{R}\times\{z_{k_m}\}}, 1_{\mathbb{R}\times\mathbb{R}\times\{z_{k'_m}\}})$.*

Lemma 4.1 provides a necessary condition based on Kitagawa (2015), Mourifié and Wan (2017), and Sun (2021) for the validity pair set $\mathscr{Z}_{\bar{M}}$. Define

$$\mathcal{G}_0 = \left\{ g \in \mathcal{G}_P : \sup_{h\in\mathcal{H}} \phi(h, g) = 0 \right\} \text{ and } \widehat{\mathcal{G}}_0 = \left\{ g \in \mathcal{G}_P : \sqrt{T_n} \left| \sup_{h\in\mathcal{H}} \frac{\widehat{\phi}(h, g)}{\xi_0 \vee \widehat{\sigma}(h, g)} \right| \leq \tau_n \right\}, \tag{4.5}$$

where $\tau_n \to \infty$ with $\tau_n/\sqrt{n} \to 0$ as $n \to \infty$, and ξ_0 is a small positive number.[11] The set $\mathcal{G}_0$ is different from the contact sets defined in Beare and Shi (2019), Sun and Beare (2021), and Sun (2021) in independent contexts because of the presence of the map sup. A further discussion about the estimation of contact sets can be found in Linton et al. (2010) and Lee et al. (2013). Define $\mathscr{Z}_0$ as the collection of all (z, z') associated with some $g \in \mathcal{G}_0$:

$$\mathscr{Z}_0 = \left\{ (z_k, z_{k'}) \in \mathscr{Z} : g = (1_{\mathbb{R}\times\mathbb{R}\times\{z_k\}}, 1_{\mathbb{R}\times\mathbb{R}\times\{z_{k'}\}}) \in \mathcal{G}_0 \right\}. \tag{4.6}$$

For example, if $K = 4$ and $\mathcal{G}_0 = \{(1_{\mathbb{R}\times\mathbb{R}\times\{z_1\}}, 1_{\mathbb{R}\times\mathbb{R}\times\{z_2\}}), (1_{\mathbb{R}\times\mathbb{R}\times\{z_3\}}, 1_{\mathbb{R}\times\mathbb{R}\times\{z_4\}})\}$, then $\mathscr{Z}_0 = \{(z_1, z_2), (z_3, z_4)\}$. By Lemma 4.1, $\mathscr{Z}_{\bar{M}} \subset \mathscr{Z}_0$. We use $\widehat{\mathcal{G}}_0$ to construct the estimator of $\mathscr{Z}_0$, denoted by $\widehat{\mathscr{Z}_0}$, which is defined as the set of all (z, z') associated with some $g \in \widehat{\mathcal{G}}_0$:

$$\widehat{\mathscr{Z}_0} = \left\{ (z_k, z_{k'}) \in \mathscr{Z} : g = (1_{\mathbb{R}\times\mathbb{R}\times\{z_k\}}, 1_{\mathbb{R}\times\mathbb{R}\times\{z_{k'}\}}) \in \widehat{\mathcal{G}}_0 \right\}. \tag{4.7}$$

Note that (4.7) is the sample analog of (4.6). The following proposition establishes consistency of $\widehat{\mathscr{Z}_0}$.

[11] In practice, we use $\xi_0 = 0.001$.

Proposition 4.1 *Under Assumption 3.1,* $\mathbb{P}(\widehat{\mathcal{G}}_0 = \mathcal{G}_0) \to 1$*, and thus* $\mathbb{P}(\widehat{\mathscr{Z}_0} = \mathscr{Z}_0) \to 1$.

Proposition 4.1 is related to the contact set estimation in Sun (2021). Since, by definition, $\mathcal{G}_0 \subset \mathcal{G}_P$ and $\mathcal{G}_P$ is a finite set, we can use techniques similar to those in Sun (2021) to obtain the stronger result in Proposition 4.1, that is, $\mathbb{P}(\widehat{\mathcal{G}}_0 = \mathcal{G}_0) \to 1$.

5 Simulation Evidence

Here we evaluate the finite sample performance of our method in Monte Carlo simulation. In Section 6, we present additional Monte Carlo evidence based on our empirical application. We consider the case where $D \in \{0,1\}$ and $Z \in \{0,1,2\}$. The presumed validity set is $\mathscr{Z}_P = \{(0,1),(0,2),(1,2)\}$. For each simulation, we use 1,000 Monte Carlo iterations. To calculate the supremum in $\sqrt{T_n}|\sup_{h\in\mathcal{H}} \widehat{\phi}\,(h,g)/(\xi_0 \vee \widehat{\sigma}(h,g))|$ for every g, we use the approach employed by Kitagawa (2015) and Sun (2021). Specifically, we compute the supremum based on the closed intervals $[a,b]$ with the realizations of $\{Y_i\}$ as endpoints, i.e., intervals $[a,b]$ where $a,b \in \{Y_i\}$ and $a \leq b$. We consider four data generating processes (DGPs) where Assumption 2.1 does not fully hold. These DGPs are constructed based on those used in Kitagawa (2015) and Sun (2021). We consider two different sample sizes $n \in \{1500, 3000\}$ and report results for $\tau_n \in \{2, 2.5, \ldots, 6.5\}$.

For all DGPs, we specify $U \sim \text{Unif}(0,1)$, $V \sim \text{Unif}(0,1)$, $W \sim \text{Unif}(0,1)$, and $Z = 2 \times 1\{U \leq 0.3\} + 1\{0.3 < U \leq 0.65\}$. For DGPs (1)–(4), we set $D_z = 1\{V \leq 0.5\}$ for $z = 0,1,2$, $D = \sum_{z=0}^{2} 1\{Z = z\} \times D_z$, $N_Z \sim N(0,1)$, $N_{00} = N_Z$, and $N_{dz} = N_Z$ for $d = 0,1$ and $z = 1,2$.

(1): $N_{10} \sim N(-0.7, 1)$, $Y = \sum_{z=0}^{2} 1\{Z = z\} \times (\sum_{d=0}^{1} 1\{D = d\} \times N_{dz})$

(2): $N_{10} \sim N(0, 1.675^2)$, $Y = \sum_{z=0}^{2} 1\{Z = z\} \times (\sum_{d=0}^{1} 1\{D = d\} \times N_{dz})$

(3): $N_{10} \sim N(0, 0.515^2)$, $Y = \sum_{z=0}^{2} 1\{Z = z\} \times (\sum_{d=0}^{1} 1\{D = d\} \times N_{dz})$

(4): $N_{10a} \sim N(-1, 0.125^2)$, $N_{10b} \sim N(-0.5, 0.125^2)$, $N_{10c} \sim N(0, 0.125^2)$, $N_{10d} \sim N(0.5, 0.125^2)$, $N_{10e} \sim N(1, 0.125^2)$, $N_{10} = 1\{W \leq 0.15\} \times N_{10a} + 1\{0.15 < W \leq 0.35\} \times N_{10b} + 1\{0.35 < W \leq 0.65\} \times N_{10c} + 1\{0.65 < W \leq 0.85\} \times N_{10d} + 1\{W > 0.85\} \times N_{10e}$, and $Y = \sum_{z=0}^{2} 1\{Z = z\} \times (\sum_{d=0}^{1} 1\{D = d\} \times N_{dz})$

The random variables U, V, W, N_Z, and N_{10} are mutually independent. Note that, for all DGPs, $\mathscr{Z}_{\bar{M}} = \mathscr{Z}_0 \cap \mathscr{Z}_P = \{(1,2)\}$. Tables 5.1–5.2 show the empirical probabilities with

which each element of $\mathscr{Z}_P$ is selected to be in $\widehat{\mathscr{Z}_0}$ in the simulations. The results show that choosing τ_n is subject to a trade-off between the ability of our method to screen-out invalid pairs and its ability to include valid pairs. Given the nature of the method, screening-out invalid pairs is particularly important since IV estimation using these pairs yields biased estimates. For $n = 1500$, choosing $\tau_n = 3.5$ allows for excluding invalid pairs with high probability across all DGPs while selecting valid pairs with relatively high probability. For $n = 3000$, our method with $\tau_n = 4$ detects invalid pairs almost perfectly while selecting valid pairs with high probability. Overall, the simulation results show that the proposed method performs well in identifying the validity pair set in finite samples.

In empirical practice, we suggest choosing τ_n using application-based Monte Carlo simulations. We illustrate this approach in Section 6.

Table 5.1: Validity Pair Set Estimation ($n = 1500$): Selection Probabilities

τ_n	DGP (1)			DGP (2)			DGP (3)			DGP (4)		
	(0, 1)	(0, 2)	**(1, 2)**	(0, 1)	(0, 2)	**(1, 2)**	(0, 1)	(0, 2)	**(1, 2)**	(0, 1)	(0, 2)	**(1, 2)**
2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2.5	0.000	0.000	0.003	0.000	0.000	0.003	0.000	0.000	0.003	0.000	0.000	0.003
3	0.000	0.001	0.209	0.000	0.000	0.209	0.000	0.000	0.209	0.002	0.001	0.209
3.5	0.000	0.002	0.754	0.000	0.001	0.754	0.001	0.002	0.754	0.049	0.052	0.754
4	0.010	0.012	0.970	0.020	0.020	0.970	0.006	0.017	0.970	0.195	0.241	0.970
4.5	0.036	0.057	0.994	0.141	0.143	0.994	0.036	0.065	0.994	0.462	0.513	0.994
5	0.109	0.155	1.000	0.410	0.406	1.000	0.113	0.141	1.000	0.721	0.765	1.000
5.5	0.256	0.308	1.000	0.718	0.720	1.000	0.243	0.279	1.000	0.888	0.917	1.000
6	0.457	0.530	1.000	0.913	0.914	1.000	0.458	0.490	1.000	0.960	0.971	1.000
6.5	0.662	0.741	1.000	0.976	0.984	1.000	0.661	0.691	1.000	0.986	0.993	1.000

Table 5.2: Validity Pair Set Estimation ($n = 3000$): Selection Probabilities

τ_n	DGP (1)			DGP (2)			DGP (3)			DGP (4)		
	(0, 1)	(0, 2)	**(1, 2)**	(0, 1)	(0, 2)	**(1, 2)**	(0, 1)	(0, 2)	**(1, 2)**	(0, 1)	(0, 2)	**(1, 2)**
2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2.5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
3	0.000	0.000	0.043	0.000	0.000	0.043	0.000	0.000	0.043	0.000	0.000	0.056
3.5	0.000	0.000	0.614	0.000	0.000	0.614	0.000	0.000	0.614	0.000	0.001	0.615
4	0.000	0.000	0.933	0.000	0.000	0.933	0.000	0.000	0.933	0.009	0.014	0.927
4.5	0.000	0.000	0.994	0.000	0.000	0.994	0.000	0.001	0.994	0.037	0.070	0.998
5	0.000	0.001	1.000	0.000	0.003	1.000	0.000	0.001	1.000	0.132	0.210	1.000
5.5	0.003	0.001	1.000	0.014	0.017	1.000	0.002	0.002	1.000	0.344	0.455	1.000
6	0.010	0.009	1.000	0.095	0.114	1.000	0.005	0.012	1.000	0.601	0.709	1.000
6.5	0.031	0.043	1.000	0.327	0.364	1.000	0.035	0.050	1.000	0.807	0.872	1.000

6 Empirical Application

6.1 Setup

We revisit the study of Angrist and Krueger (1991) and examine the use of the classical quarter of birth (QOB) instrument for estimating the returns to schooling. As explained by Dahl et al. (2017), the validity of this instrument has been contested. For example, Bound et al. (1995) argue that the exclusion restriction (Assumption 2.1.(i)) is not plausible because of seasonal birth patterns; see also Buckles and Hungerman (2013). Moreover, the validity of the monotonicity assumption (Assumption 2.1.(iii)) is questionable due to strategic parent behavior when enrolling their children (e.g., Barua and Lang, 2016).

Here we use the proposed method to remove invalid variation in the QOB instrument. The data set is from Angrist and Krueger (1991), and we use the same sample of 486,926 men born between 1940 and 1949 as in Dahl et al. (2017).[12] Following Dahl et al. (2017), the outcome Y is the log weekly wage, and the binary treatment D is equal to 1 if an individual has 13 or more of years of schooling and 0 otherwise. The QOB instrument $Z \in \{1,2,3,4\}$ indicates the quarter in which an individual is born. We assume that $\mathscr{Z}_P = \{(1,2),(1,3),(1,4),(2,3),(2,4),(3,4)\}$.

6.2 Choosing τ_n using Application-based Simulations

We determine the choice of τ_n using an application-based Monte Carlo simulation. We construct four DGPs similar to those in Section 5 and calibrated to match joint distribution of (D,Z) in the data. Let $U \sim \text{Unif}(0,1)$, $V \sim \text{Unif}(0,1)$, $W \sim \text{Unif}(0,1)$, $Z = 1\{U \leq 0.2418\} + 2 \times 1\{0.2418 < U \leq 0.4774\} + 3 \times 1\{0.4774 < U \leq 0.7440\} + 4 \times 1\{U > 0.7440\}$, $D_1 = 1\{V \leq 0.5104\}$, $D_2 = 1\{V \leq 0.5187\}$, $D_3 = 1\{V \leq 0.5203\}$, $D_4 = 1\{V \leq 0.5295\}$, $D = \sum_{z=1}^{4} 1\{Z = z\} \times D_z$, $N_Z \sim N(0,1)$, $N_{02} = N_Z$, and $N_{dz} = N_Z$ for $d = 0,1$ and $z = 1,3,4$. The DGPs are specified as follows.

(1): $N_{12} \sim N(-0.07, 1)$, $Y = \sum_{z=1}^{4} 1\{Z = z\} \times (\sum_{d=0}^{1} 1\{D = d\} \times N_{dz})$

(2): $N_{12} \sim N(0, 1.0675^2)$, $Y = \sum_{z=1}^{4} 1\{Z = z\} \times (\sum_{d=0}^{1} 1\{D = d\} \times N_{dz})$

(3): $N_{12} \sim N(0, 0.9325^2)$, $Y = \sum_{z=1}^{4} 1\{Z = z\} \times (\sum_{d=0}^{1} 1\{D = d\} \times N_{dz})$

[12] The data set was downloaded from https://economics.mit.edu/faculty/angrist/data1/data/angkru1991 (last accessed February 5, 2022).

(4): $N_{12a} \sim N(-0.1, 0.925^2)$, $N_{12b} \sim N(-0.05, 0.925^2)$, $N_{12c} \sim N(0, 0.925^2)$, $N_{12d} \sim N(0.05, 0.925^2)$, $N_{12e} \sim N(0.1, 0.925^2)$, $N_{12} = 1\{W \leq 0.15\} \times N_{12a} + 1\{0.15 < W \leq 0.35\} \times N_{12b} + 1\{0.35 < W \leq 0.65\} \times N_{12c} + 1\{0.65 < W \leq 0.85\} \times N_{12d} + 1\{W > 0.85\} \times N_{12e}$, and $Y = \sum_{z=1}^{4} 1\{Z = z\} \times (\sum_{d=0}^{1} 1\{D = d\} \times N_{dz})$

The random variables U, V, W, N_Z, and N_{12} are mutually independent. For all DGPs, $\mathscr{Z}_{\bar{M}} = \mathscr{Z}_0 \cap \mathscr{Z}_P = \{(1,3), (1,4), (3,4)\}$. These four DGPs match the empirical proportions for $Z = 1$, $Z = 2$, $Z = 3$, and $Z = 4$ (0.2418, 0.2356, 0.2666, and 0.2560, respectively) as well as the proportions for $D = 1$ given $Z = 1$, $D = 1$ given $Z = 2$, $D = 1$ given $Z = 3$, and $D = 1$ given $Z = 4$ (0.5104, 0.5187, 0.5203, and 0.5295, respectively).

Since the sample size is very large, for computational tractability, we randomly choose 200 observations from $\{Y_i\}$ to construct closed intervals for $\mathcal{H}$. We report simulation results based on 1,000 repetitions. For each repetition r, we denote the $\mathcal{H}$ constructed by the 200 randomly chosen observations by $\mathcal{H}_r$, and we use $\mathcal{H}_r$ to construct

$$\sqrt{T_n} \max \left\{ \sup_{h \in \mathcal{H}_r} \widehat{\phi}\,(h,g)/(\xi_0 \vee \widehat{\sigma}(h,g)), 0 \right\}.$$

Note that for every r,

$$\sqrt{T_n} \max \left\{ \sup_{h \in \mathcal{H}_r} \widehat{\phi}\,(h,g)/(\xi_0 \vee \widehat{\sigma}(h,g)), 0 \right\} \leq \sqrt{T_n} \left| \sup_{h \in \mathcal{H}} \widehat{\phi}\,(h,g)/(\xi_0 \vee \widehat{\sigma}(h,g)) \right|. \tag{6.1}$$

We use $\sqrt{T_n} \max\{\sup_{h \in \mathcal{H}_r} \widehat{\phi}\,(h,g)/(\xi_0 \vee \widehat{\sigma}(h,g)), 0\}$ in each iteration of the simulations for the estimation of the validity pair set. If $\mathcal{H}_r = \mathcal{H}$, the equality in (6.1) holds. As $\mathcal{H}_r$ increases to $\mathcal{H}$, the simulation results would converge to those from using the statistic $\sqrt{T_n} |\sup_{h \in \mathcal{H}} \widehat{\phi}\,(h,g)/(\xi_0 \vee \widehat{\sigma}(h,g))|$. Table 6.1 shows the empirical inclusion probabilities of all pairs in $\mathscr{Z}_P$. Given the nature of our approach, which is based on necessary (but not necessarily sufficient) conditions, we recommend choosing τ_n conservatively: We prefer smaller values of τ_n, provided that the selection rates for valid pairs are high enough. When τ_n is larger than or equal to 4, the selection rates for the valid pairs $(1,3)$, $(1,4)$, and $(3,4)$ are all close to 100%, while the selection rates for the invalid pairs $(1,2)$, $(2,3)$, and $(2,4)$ are still below 0.5% across all DGPs for $\tau_n \leq 4.1$. This suggests that a reasonable conservative choice is $\tau_n = 4$. For this choice, our method screens out invalid pairs with high probability while maintaining high selection rates for the valid pairs.

Table 6.1: Validity Pair Set Estimation Application-based DGPs: Selection Probabilities

τ_n	DGP (1)						DGP (2)					
	(1, 2)	**(1, 3)**	**(1, 4)**	(2, 3)	(2, 4)	**(3, 4)**	(1, 2)	**(1, 3)**	**(1, 4)**	(2, 3)	(2, 4)	**(3, 4)**
1.5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	0.000	0.001	0.010	0.000	0.000	0.000	0.000	0.001	0.016	0.000	0.000	0.001
2.5	0.000	0.109	0.361	0.000	0.000	0.108	0.000	0.104	0.347	0.000	0.000	0.104
3	0.000	0.622	0.832	0.000	0.000	0.617	0.000	0.624	0.828	0.000	0.000	0.613
3.5	0.000	0.932	0.976	0.000	0.000	0.923	0.001	0.936	0.976	0.000	0.000	0.924
4	0.000	0.991	0.997	0.000	0.001	0.993	0.003	0.991	0.998	0.000	0.000	0.990
4.1	0.000	0.995	0.998	0.000	0.001	0.995	0.004	0.994	0.999	0.000	0.000	0.993
4.2	0.000	0.997	0.999	0.000	0.001	1.000	0.006	0.996	1.000	0.000	0.000	0.998
4.3	0.000	0.997	0.999	0.000	0.002	1.000	0.008	0.996	1.000	0.000	0.000	0.998
4.4	0.000	1.000	1.000	0.000	0.003	1.000	0.011	0.998	1.000	0.000	0.000	1.000
4.5	0.001	1.000	1.000	0.000	0.004	1.000	0.020	0.999	1.000	0.000	0.000	1.000
5	0.012	1.000	1.000	0.000	0.024	1.000	0.064	1.000	1.000	0.000	0.004	1.000
τ_n	DGP (3)						DGP (4)					
	(1, 2)	**(1, 3)**	**(1, 4)**	(2, 3)	(2, 4)	**(3, 4)**	(1, 2)	**(1, 3)**	**(1, 4)**	(2, 3)	(2, 4)	**(3, 4)**
1.5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	0.000	0.001	0.017	0.000	0.000	0.001	0.000	0.002	0.014	0.000	0.000	0.002
2.5	0.000	0.109	0.359	0.000	0.000	0.113	0.000	0.109	0.366	0.000	0.000	0.106
3	0.000	0.621	0.827	0.000	0.000	0.614	0.000	0.628	0.838	0.000	0.000	0.612
3.5	0.000	0.932	0.976	0.000	0.000	0.922	0.000	0.929	0.979	0.000	0.000	0.914
4	0.000	0.989	0.996	0.000	0.000	0.991	0.000	0.991	0.997	0.000	0.001	0.991
4.1	0.000	0.995	0.997	0.000	0.001	0.993	0.000	0.994	0.999	0.000	0.002	0.994
4.2	0.000	0.997	0.999	0.000	0.003	0.999	0.000	0.997	1.000	0.000	0.002	0.999
4.3	0.000	0.998	1.000	0.000	0.006	1.000	0.000	0.997	1.000	0.000	0.002	1.000
4.4	0.000	1.000	1.000	0.000	0.012	1.000	0.000	0.999	1.000	0.000	0.002	1.000
4.5	0.000	1.000	1.000	0.000	0.017	1.000	0.000	1.000	1.000	0.000	0.002	1.000
5	0.000	1.000	1.000	0.000	0.054	1.000	0.000	1.000	1.000	0.000	0.008	1.000

6.3 Empirical Results

To ensure computational tractability, we construct $\mathcal{H}$ based on a random subsample of 200 observations as in Section 6.2. Specifically, we estimate $\mathscr{Z}_{\bar{M}}$ by the intersection of $\mathscr{Z}_P$ and $\widehat{\mathscr{Z}_0}$ in (4.7) with

$$\widehat{\mathcal{G}_0} = \left\{ g \in \mathcal{G}_P : \sqrt{T_n} \max \left\{ \sup_{h \in \mathcal{H}} \widehat{\phi}\,(h, g)/(\xi_0 \vee \widehat{\sigma}(h, g)), 0 \right\} \leq \tau_n \right\}, \tag{6.2}$$

where $\mathcal{H}$ is constructed as described above.

The results in Table 6.2 show that as τ_n increases, the number of selected pairs increases (as expected). The simulations in Section 6.2 show that $\tau_n = 4.0$ is a reasonable conservative tuning parameter choice. For this choice, all pairs except for $(2, 4)$ are selected. The pair $(2, 4)$ remains excluded for $\tau_n \leq 4.3$. This suggests that one should be careful about using the instrument value pair $(2, 4)$ (the contrast between the second and

the fourth quarter) in this application.

The last row of Table 6.2 presents the VSIV estimates $\widehat{\beta}^1_{(k,k')}$ for $\tau_n = 4.0$. If $(2,4)$ were included in the estimated validity pair set, then the corresponding LATE would be negative, $\widehat{\beta}^1_{(2,4)} = -0.9049$. It is interesting to note that $\widehat{\beta}^1_{(1,4)}$ and $\widehat{\beta}^1_{(3,4)}$ are negative.[13] One possible explanation for this counterintuitive finding is that pairwise IV validity does not hold for the pairs $(1,4)$ and $(3,4)$, consistent with the concerns about the validity of the QOB instrument discussed above.[14] VSIV estimation detects the invalid pair $(2,4)$ (for which the effect is also negative), but because it relies on necessary conditions and because we only use 200 observations to construct $\mathcal{H}$ to ensure computational tractability given the very large sample size, it may not detect all invalid pairs.

The results in this empirical application demonstrate that VSIV estimation is a valuable tool for screening out invalid variation in the QOB instrument. Even if the information in the testable restrictions for IV validity is not sufficient for screening out all invalid pairs, VSIV estimation reduces the bias relative to standard IV methods.

Table 6.2: Validity Pair Set Estimation in Application

τ_n	(1, 2)	(1, 3)	(1, 4)	(2, 3)	(2, 4)	(3, 4)
2	0	0	0	0	0	0
2.5	0	1	0	0	0	0
3	1	1	0	0	0	1
3.5	1	1	1	0	0	1
4	1	1	1	1	0	1
4.1	1	1	1	1	0	1
4.2	1	1	1	1	0	1
4.3	1	1	1	1	0	1
4.4	1	1	1	1	1	1
4.5	1	1	1	1	1	1
$\widehat{\beta}^1_{(k,k')}$	0.2870	0.2706	-0.3858	0.1836	0	-1.0902

[13] The negative effects are due to negative reduced-form estimates.

[14] Dahl et al. (2017) exclude the winter quarters altogether due to concerns related to winter births being disproportionately by teenagers and unmarried women.

References

Angrist, J. D. and Evans, W. N. (1998). Children and their parents' labor supply: Evidence from exogenous variation in family size. *American Economic Review*, 88(3):450–477.

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 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. and Kolesár, M. (2021). Sensitivity analysis using approximate moment condition models. *Quantitative Economics*, 12(1):77–108.

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.

Barua, R. and Lang, K. (2016). School entry, educational attainment, and quarter of birth: A cautionary tale of a local average treatment effect. *Journal of Human Capital*, 10(3):347–376.

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.

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

Chernozhukov, V. and Hansen, C. (2005). An IV model of quantile treatment effects. *Econometrica*, 73(1):245–261.

Conley, T. G., Hansen, C. B., and Rossi, P. E. (2012). Plausibly Exogenous. *The Review of Economics and Statistics*, 94(1):260–272.

Dahl, C. M., Huber, M., and Mellace, G. (2017). It's never too late: A new look at local average treatment effects with or without defiers. Working Paper.

Farbmacher, H., Guber, R., and Klaassen, S. (2022). Instrument validity tests with causal forests. *Journal of Business & Economic Statistics*, 40(2):605–614.

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

Frölich, M. (2007). Nonparametric IV estimation of local average treatment effects with covariates. *Journal of Econometrics*, 139(1):35–75. Endogeneity, instruments and identification.

Fusejima, K. (2020). Identification of multi-valued treatment effects with unobserved heterogeneity. arXiv:2010.04385.

Goff, L. (2020). A vector monotonicity assumption for multiple instruments. arXiv:2009.00553.

Goh, G. and Yu, J. (2022). Causal inference with some invalid instrumental variables: A quasi-bayesian approach. *Oxford Bulletin of Economics and Statistics*, n/a(n/a).

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

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

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 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., Li, L., and Mourifié, I. (2020). Discordant relaxations of misspecified models. arXiv:2012.11679.

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.

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.

Masten, M. A. and Poirier, A. (2021). Salvaging falsified instrumental variable models. *Econometrica*, 89(3):1449–1469.

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. (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.

Sun, Z. (2021). Instrument validity for heterogeneous causal effects. arXiv:2009.01995.

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

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

Appendix to *Pairwise Valid Instruments*

Zhenting Sun Kaspar Wüthrich

The Appendix consists of three sections. Section A extends the results in the main text to multivalued ordered and unordered treatments. Section B provides the proofs and supplementary results for Section 2 and Appendix A.1. Section C provides the proofs and supplementary results for Appendix A.2.

A Extension: Multivalued Ordered and Unordered Treatments

In this section, we generalize the results in the main text to multivalued ordered and unordered treatments.

A.1 Ordered Treatments

Suppose, in general, that the observable treatment variable $D \in \mathcal{D} = \{d_1, \ldots, d_J\}$. Without loss of generality, suppose $d_1 < \cdots < d_J$. The following assumption is a straightforward generalization of Assumption 2.1 to ordered treatments (e.g., Sun, 2021).

Assumption A.1 *IV Validity Conditions for Ordered Treatments:*

(i) Exclusion: For all $d \in \mathcal{D}$, $Y_{dz_1} = Y_{dz_2} = \cdots = Y_{dz_K}$ *a.s.*

(ii) Random Assignment: Z *is jointly independent of* $(Y_{d_1 z_1}, \ldots, Y_{d_1 z_K}, \ldots, Y_{d_J z_1}, \ldots, Y_{d_J z_K})$ *and* $(D_{z_1}, \ldots, D_{z_K})$.

(iii) Monotonicity: For all $k = 1, \ldots, K-1$, $D_{z_{k+1}} \geq D_{z_k}$ *a.s.*

We next introduce the definition of pairwise valid instruments for ordered treatments.

Definition A.1 *The instrument* Z *is* ***pairwise valid*** *for the ordered treatment* $D \in \mathcal{D} = \{d_1, \ldots, d_J\}$ *if there is a set* $\mathscr{Z}_M = \{(z_{k_1}, z_{k'_1}), \ldots, (z_{k_M}, z_{k'_M})\}$ *with* $z_{k_1}, z_{k'_1}, \ldots, z_{k_M}, z_{k'_M} \in \mathcal{Z}$ *such that the following conditions hold for every* $(z, z') \in \mathscr{Z}_M$*:*

(i) Exclusion: For all $d \in \mathcal{D}$, $Y_{dz} = Y_{dz'}$ a.s.

(ii) Random Assignment: Z is jointly independent of $(Y_{d_1 z}, Y_{d_1 z'}, \ldots, Y_{d_J z}, Y_{d_J z'}, D_z, D_{z'})$.

(iii) Monotonicity: $D_{z'} \geq D_z$ a.s.

*The set $\mathscr{Z}_M$ is called a **validity pair set** of Z. The union of all validity pair sets is the largest validity pair set, denoted by $\mathscr{Z}_{\bar{M}}$.*

With the exclusion condition, for every $(z, z') \in \mathscr{Z}_{\bar{M}}$, define $Y_d(z, z')$ such that $Y_d(z, z') = Y_{dz} = Y_{dz'}$ a.s. for all $d \in \mathcal{D}$.

Lemma A.1 *Suppose that the instrument Z is pairwise valid as defined in Definition A.1 with a known validity pair set $\mathscr{Z}_M = \{(z_{k_1}, z_{k'_1}), \ldots, (z_{k_M}, z_{k'_M})\}$. Then for every $m \in \{1, \ldots, M\}$, the following quantity can be identified:*

$$\begin{aligned} \beta_{k'_m, k_m} &\equiv \sum_{j=2}^{J} \omega_j \cdot E\left[\left(Y_{d_j}(z_{k_m}, z_{k'_m}) - Y_{d_{j-1}}(z_{k_m}, z_{k'_m})\right) | D_{z_{k'_m}} \geq d_j > D_{z_{k_m}}\right] \\ &= \frac{E\left[Y | Z = z_{k'_m}\right] - E\left[Y | Z = z_{k_m}\right]}{E\left[D | Z = z_{k'_m}\right] - E\left[D | Z = z_{k_m}\right]}, \end{aligned} \tag{A.1}$$

where

$$\omega_j = \frac{\mathbb{P}\left(D_{z_{k'_m}} \geq d_j > D_{z_{k_m}}\right)}{\sum_{l=2}^{J} (d_l - d_{l-1}) \mathbb{P}\left(D_{z_{k'_m}} \geq d_l > D_{z_{k_m}}\right)}.$$

Lemma A.1 is an extension of Theorem 1 of Imbens and Angrist (1994) and Theorem 1 of Angrist and Imbens (1995) to the case where Z is pairwise valid. We follow Angrist and Imbens (1995) and refer to $\beta_{k'_m, k_m}$ as the average causal response (ACR). Lemma A.1 shows that if a validity pair set $\mathscr{Z}_M$ is known, we can identify every $\beta_{k'_m, k_m}$. In practice, however, $\mathscr{Z}_M$ is usually unknown. We show how to identify the largest validity pair set $\mathscr{Z}_{\bar{M}}$ and use it to estimate the ACRs.

The estimation of $\mathscr{Z}_{\bar{M}}$ is similar to that in Section 2. Suppose that there are subsets $\mathscr{Z}_1 \subset \mathscr{Z}$ and $\mathscr{Z}_2 \subset \mathscr{Z}$ that satisfy the testable implications in Kitagawa (2015), Mourifié and Wan (2017), and Sun (2021), and those in Kédagni and Mourifié (2020), respectively. We let $\mathscr{Z}_0 = \mathscr{Z}_1 \cap \mathscr{Z}_2$ so that $\mathscr{Z}_0$ satisfies all the above necessary conditions. We first construct the estimators $\widehat{\mathscr{Z}_1}$ and $\widehat{\mathscr{Z}_2}$ for $\mathscr{Z}_1$ and $\mathscr{Z}_2$, respectively, and then construct the estimator $\widehat{\mathscr{Z}_0}$ for $\mathscr{Z}_0$ as $\widehat{\mathscr{Z}_0} = \widehat{\mathscr{Z}_1} \cap \widehat{\mathscr{Z}_2}$. See Appendix B.4 for details.

Assumption A.2 *$\{(Y_i, D_i, Z_i)\}_{i=1}^n$ is an i.i.d. sample from a population such that all relevant moments exist.*

Assumption A.3 *For every $\mathcal{Z}_{(k,k')} \in \mathscr{Z}_{\bar{M}}$,*

$$E[g(Z_i)D_i|Z_i \in \mathcal{Z}_{(k,k')}] - E[D_i|Z_i \in \mathcal{Z}_{(k,k')}] \cdot E[g(Z_i)|Z_i \in \mathcal{Z}_{(k,k')}] \neq 0. \tag{A.2}$$

As in Section 2, we first suppose that $\mathscr{Z}_{\bar{M}}$ can be estimated consistently by the estimator $\widehat{\mathscr{Z}_0}$. We use the same notation as in Section 2. For $\mathcal{Z}_{(k,k')} \in \mathscr{Z}$, we run the regression

$$Y_i 1\left\{Z_i \in \mathcal{Z}_{(k,k')}\right\} = \gamma^0_{(k,k')} 1\left\{Z_i \in \mathcal{Z}_{(k,k')}\right\} + \gamma^1_{(k,k')} D_i 1\left\{Z_i \in \mathcal{Z}_{(k,k')}\right\} + \epsilon_i 1\left\{Z_i \in \mathcal{Z}_{(k,k')}\right\}, \tag{A.3}$$

using $g(Z_i)1\{Z_i \in \mathcal{Z}_{(k,k')}\}$ as the instrument for $D_i 1\{Z_i \in \mathcal{Z}_{(k,k')}\}$. Given the estimated validity set $\widehat{\mathscr{Z}_0}$, we define the VSIV estimator for each $\mathcal{Z}_{(k,k')}$ as

$$\widehat{\beta}^1_{(k,k')} = 1\left\{\mathcal{Z}_{(k,k')} \in \widehat{\mathscr{Z}_0}\right\} \cdot \frac{\mathcal{E}_n\left(g\left(Z_i\right) Y_i, \mathcal{Z}_{(k,k')}\right) - \mathcal{E}_n\left(g\left(Z_i\right), \mathcal{Z}_{(k,k')}\right) \mathcal{E}_n\left(Y_i, \mathcal{Z}_{(k,k')}\right)}{\mathcal{E}_n\left(g\left(Z_i\right) D_i, \mathcal{Z}_{(k,k')}\right) - \mathcal{E}_n\left(g\left(Z_i\right), \mathcal{Z}_{(k,k')}\right) \mathcal{E}_n\left(D_i, \mathcal{Z}_{(k,k')}\right)}, \tag{A.4}$$

which is the IV estimator for $\gamma^1_{(k,k')}$ in (A.3) multiplied by $1\{\mathcal{Z}_{(k,k')} \in \widehat{\mathscr{Z}_0}\}$. As in Section 2, we define

$$\widehat{\beta}_1 = \left(\widehat{\beta}^1_{(1,2)}, \ldots, \widehat{\beta}^1_{(1,K)}, \ldots, \widehat{\beta}^1_{(K,1)}, \ldots, \widehat{\beta}^1_{(K,K-1)}\right)^T,$$

$$\beta^1_{(k,k')} = 1\left\{\mathcal{Z}_{(k,k')} \in \mathscr{Z}_{\bar{M}}\right\} \cdot \frac{\mathcal{E}\left(g\left(Z_i\right) Y_i, \mathcal{Z}_{(k,k')}\right) - \mathcal{E}\left(g\left(Z_i\right), \mathcal{Z}_{(k,k')}\right) \mathcal{E}\left(Y_i, \mathcal{Z}_{(k,k')}\right)}{\mathcal{E}\left(g\left(Z_i\right) D_i, \mathcal{Z}_{(k,k')}\right) - \mathcal{E}\left(g\left(Z_i\right), \mathcal{Z}_{(k,k')}\right) \mathcal{E}\left(D_i, \mathcal{Z}_{(k,k')}\right)}, \tag{A.5}$$

and

$$\beta_1 = \left(\beta^1_{(1,2)}, \ldots, \beta^1_{(1,K)}, \ldots, \beta^1_{(K,1)}, \ldots, \beta^1_{(K,K-1)}\right)^T.$$

Theorem A.1 *Suppose that the instrument Z is pairwise valid for the treatment D as defined in Definition A.1 with the largest validity pair set $\mathscr{Z}_{\bar{M}} = \{(z_{k_1}, z_{k'_1}), \ldots, (z_{k_{\bar{M}}}, z_{k'_{\bar{M}}})\}$ and that the estimator $\widehat{\mathscr{Z}_0}$ satisfies $\mathbb{P}(\widehat{\mathscr{Z}_0} = \mathscr{Z}_{\bar{M}}) \to 1$. Under Assumptions A.2 and A.3, $\sqrt{n}(\widehat{\beta}_1 - \beta_1) \xrightarrow{d} N(0, \Sigma)$, where Σ is defined in (B.5). In addition, $\beta^1_{(k,k')} = \beta_{k',k}$ as defined in (A.1) for every $(z_k, z_{k'}) \in \mathscr{Z}_{\bar{M}}$.*

Next, we generalize the results in Section 3.2 and show that VSIV estimation always reduces the asymptotic estimation bias when the treatments are ordered. Given a presumed validity pair set $\mathscr{Z}_P$, we apply VSIV estimation based on $\widehat{\mathscr{Z}'_0} = \widehat{\mathscr{Z}_0} \cap \mathscr{Z}_P$.

Assumption A.4 *For every* $\mathcal{Z}_{(k,k')} \in \mathscr{Z}_0$*,*

$$E[g(Z_i)D_i|Z_i \in \mathcal{Z}_{(k,k')}] - E[D_i|Z_i \in \mathcal{Z}_{(k,k')}] \cdot E[g(Z_i)|Z_i \in \mathcal{Z}_{(k,k')}] \neq 0. \tag{A.6}$$

Theorem A.2 *Suppose that Assumptions A.2 and A.4 hold and that* $\mathbb{P}(\widehat{\mathscr{Z}_0} = \mathscr{Z}_0) \to 1$ *with* $\mathscr{Z}_0 \supset \mathscr{Z}_{\bar{M}}$*. For every presumed validity pair set* $\mathscr{Z}_P$*, the asymptotic estimation bias* $\operatorname{plim}_{n\to\infty}\|\widehat{\beta}_1 - \beta_1\|_2$ *is always reduced by using* $\widehat{\mathscr{Z}_0'}$ *in the estimation* (A.4) *compared to the bias from using* $\mathscr{Z}_P$*.*

As shown in Propositions B.1 and B.2, the pseudo-validity pair set $\mathscr{Z}_0$ can always be estimated consistently by $\widehat{\mathscr{Z}_0}$ under mild conditions. Theorem A.2 shows that VSIV estimation based on $\widehat{\mathscr{Z}_0} \cap \mathscr{Z}_P$ always reduces the bias.

Remark A.1 *In Section 2, we provide the definition of partial IV validity for the binary treatment case. See Appendix B.5 for the extension to multivalued ordered treatments.*

A.2 Unordered Treatments

A.2.1 Setup

Here, we extend our results to unordered treatments using the framework of Heckman and Pinto (2018). The treatment (choice) D is discrete with support $\mathcal{D} = \{d_1, \ldots, d_J\}$, which is unordered. Heckman and Pinto (2018, p. 15) (Assumption A-3) consider the following monotonicity assumption.

Assumption A.5 *For all* $d \in \mathcal{D}$ *and all* $z, z' \in \mathcal{Z}$*,* $1\{D_{z'} = d\} \geq 1\{D_z = d\}$ *for all* $\omega \in \Omega$*, or* $1\{D_{z'} = d\} \leq 1\{D_z = d\}$ *for all* $\omega \in \Omega$*.*[15]

Based on Assumption A.5, we introduce the definition of the pairwise IV validity for the unordered treatment case.[16]

Definition A.2 *The instrument* Z *is* ***pairwise valid*** *for the unordered treatment* D *if there is a set* $\mathscr{Z}_M = \{(z_{k_1}, z_{k_1'}), \ldots, (z_{k_M}, z_{k_M'})\}$ *with* $z_{k_1}, z_{k_1'}, \ldots, z_{k_M}, z_{k_M'} \in \mathcal{Z}$ *and* $k_m < k_m'$ *for every* m *such that the following conditions hold for every* $(z, z') \in \mathscr{Z}_M$*:*

[15] More precisely, the potential treatments should be written as functions of ω, $D_z(\omega)$ and $D_{z'}(\omega)$. For simplicity of notation, we omit ω whenever there is no confusion. The inequalities can be modified to hold a.s.

[16] Fusejima (2020) combines a similar assumption with rank similarity (Chernozhukov and Hansen, 2005) to identify effects with multivalued treatments.

(i) Exclusion: For all $d \in \mathcal{D}$, $Y_{dz} = Y_{dz'}$ *a.s.*

(ii) Random Assignment: Z *is jointly independent of* $(Y_{d_1 z}, Y_{d_1 z'}, \ldots, Y_{d_J z}, Y_{d_J z'}, D_z, D_{z'})$.

(iii) Monotonicity: For all $d \in \mathcal{D}$, $1\{D_{z'} = d\} \geq 1\{D_z = d\}$ *for all* $\omega \in \Omega$, *or* $1\{D_{z'} = d\} \leq 1\{D_z = d\}$ *for all* $\omega \in \Omega$.

The set $\mathscr{Z}_M$ *is called a* ***validity pair set*** *of* Z. *The union of all validity pair sets is the largest validity pair set, denoted by* $\mathscr{Z}_{\bar{M}}$.

Suppose the instrument Z is pairwise valid for the treatment D with the largest validity pair set $\mathscr{Z}_{\bar{M}} = \{(z_{k_1}, z_{k'_1}), \ldots, (z_{k_{\bar{M}}}, z_{k'_{\bar{M}}})\}$. Define $Y_d(z, z')$ for every $d \in \mathcal{D}$ and every $(z, z') \in \mathscr{Z}_{\bar{M}}$ such that $Y_d(z, z') = Y_{dz} = Y_{dz'}$ a.s. Following Heckman and Pinto (2018), we introduce the following notation. Define the response vector S as a K-dimensional random vector of potential treatments with Z fixed at each value of its support:

$$S = (D_{z_1}, \ldots, D_{z_K})^T .$$

The finite support of S is $\mathcal{S} = \{\xi_1, \ldots, \xi_{N_S}\}$, where N_S is the number of possible values of S. The response matrix R is an array of response-types defined over $\mathcal{S}$, $R = (\xi_1, \ldots, \xi_{N_S})$.

For every $\mathcal{Z}_{(k,k')} \in \mathscr{Z}$, there is a $2 \times K$ binary matrix $\mathcal{M}_{(k,k')}$ such that

$$\mathcal{M}_{(k,k')} (z_1, \ldots, z_K)^T = (z_k, z_{k'})^T .$$

For example, if $K = 5$ and $(k, k') = (3, 5)$, then

$$\mathcal{M}_{(3,5)} = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

We define a transformation $\mathcal{K}_{(k,k')}$ such that if A is a $K \times L$ matrix, $\mathcal{K}_{(k,k')}A$ is the matrix that consists of all the unique columns of $\mathcal{M}_{(k,k')}A$ in the same order as in $\mathcal{M}_{(k,k')}A$. In the above example, if $A = ((x_1, \ldots, x_5)^T, (x_1, \ldots, x_5)^T, (y_1, \ldots, y_5)^T)$, then $\mathcal{K}_{(3,5)}A = ((x_3, x_5)^T, (y_3, y_5)^T)$. We write $\mathcal{K}_{(k,k')}R = (s_1, \ldots, s_{L_{(k,k')}})$, where $L_{(k,k')}$ is the column number of $\mathcal{K}_{(k,k')}R$. Let $B_{d(k,k')}$ denote a binary matrix of the same dimension as $\mathcal{K}_{(k,k')}R$, whose elements are equal to 1 if the corresponding element in $\mathcal{K}_{(k,k')}R$ is equal to d, and equal to 0 otherwise. We denote the element in the mth row and lth column of the matrix $B_{d(k,k')}$ by $B_{d(k,k')}(m, l)$. Finally, we use $B_{d(k,k')} = 1\{\mathcal{K}_{(k,k')}R = d\}$ to denote $B_{d(k,k')}$.

Lemma A.2 *Suppose that the instrument Z is pairwise valid for the treatment D with the largest validity pair set $\mathscr{Z}_{\bar{M}} = \{(z_{k_1}, z_{k'_1}), \ldots, (z_{k_{\bar{M}}}, z_{k'_{\bar{M}}})\}$. The following statements are equivalent:*

(i) For every $(z_k, z_{k'}) \in \mathscr{Z}_{\bar{M}}$, the binary matrix $B_{d(k,k')} = 1\{\mathcal{K}_{(k,k')}R = d\}$ is lonesum[17] for every $d \in \mathcal{D}$.

(ii) For every $(z_k, z_{k'}) \in \mathscr{Z}_{\bar{M}}$ and all $d, d', d'' \in \mathcal{D}$, there are no 2×2 sub-matrices of $\mathcal{K}_{(k,k')}R$ of the type

$$\begin{pmatrix} d & d' \\ d'' & d \end{pmatrix} \text{ or } \begin{pmatrix} d' & d \\ d & d'' \end{pmatrix}$$

with $d' \neq d$ and $d'' \neq d$.

(iii) For every $(z_k, z_{k'}) \in \mathscr{Z}_{\bar{M}}$ and every $d \in \mathcal{D}$, the following inequalities hold:

$$1\left\{D_{z_{k'}} = d\right\} \geq 1\left\{D_{z_k} = d\right\} \text{ for all } \omega \in \Omega,$$
$$\text{or } 1\left\{D_{z_{k'}} = d\right\} \leq 1\left\{D_{z_k} = d\right\} \text{ for all } \omega \in \Omega.$$

Lemma A.2 is an extension of Theorem T-3 of Heckman and Pinto (2018) for pairwise valid instruments. It provides equivalent conditions for the monotonicity condition (iii) in Definition A.2.

To describe our results, following Heckman and Pinto (2018), we define some additional notation. Let $B^{+}_{d(k,k')}$ denote the Moore–Penrose pseudo-inverse of $B_{d(k,k')}$. Let $\kappa : \mathbb{R} \to \mathbb{R}$ be an arbitrary function of interest. Define for all $d \in \mathcal{D}$,

$$\bar{P}_Z(d) = \left(\mathbb{P}(D = d|Z = z_1), \ldots, \mathbb{P}(D = d|Z = z_K)\right)^T,$$

$$\bar{Q}_Z(d) = \left(E[\kappa(Y) \cdot 1\{D = d\}|Z = z_1], \ldots, E[\kappa(Y) \cdot 1\{D = d\}|Z = z_K]\right)^T,$$

$$P_{Z(k,k')}(d) = \mathcal{M}_{(k,k')}\bar{P}_Z(d) = \left(\mathbb{P}(D = d|Z = z_k), \mathbb{P}(D = d|Z = z_{k'})\right)^T,$$

and

$$\begin{aligned} Q_{Z(k,k')}(d) &= \mathcal{M}_{(k,k')}\bar{Q}_Z(d) \\ &= \left(E[\kappa(Y) \cdot 1\{D = d\}|Z = z_k], E[\kappa(Y) \cdot 1\{D = d\}|Z = z_{k'}]\right)^T. \end{aligned}$$

[17] "A binary matrix is *lonesum* if it is uniquely determined by its row and column sums." (Heckman and Pinto, 2018, p. 20)

Moreover, we define

$$P_{Z(k,k')} = \left(P_{Z(k,k')}(d_1), \ldots, P_{Z(k,k')}(d_J)\right)^T \text{ and}$$
$$P_{S(k,k')} = \left(\mathbb{P}(\mathcal{M}_{(k,k')}S = s_1), \ldots, \mathbb{P}(\mathcal{M}_{(k,k')}S = s_{L_{(k,k')}})\right)^T,$$

and for every $(z_k, z_{k'}) \in \mathscr{Z}_{\bar{M}}$, we define

$$Q_{S(k,k')}(d) =$$
$$\left(E\left[\kappa\left(Y_d(z_k, z_{k'})\right) \cdot 1\left\{\mathcal{M}_{(k,k')}S = s_1\right\}\right], \ldots, E\left[\kappa\left(Y_d(z_k, z_{k'})\right) \cdot 1\left\{\mathcal{M}_{(k,k')}S = s_{L_{(k,k')}}\right\}\right]\right)^T$$

for all $d \in \mathcal{D}$. Define $\Sigma_{d(k,k')}(t)$ to be the set of response-types in which d appears exactly t times, that is, for every $d \in \mathcal{D}$ and every $t \in \{0, 1, 2\}$, define

$$\Sigma_{d(k,k')}(t) = \left\{ s : s \text{ is some } l\text{th column of } \mathcal{K}_{(k,k')}R \text{ with } \sum_{m=1}^{2} B_{d(k,k')}(m, l) = t \right\}.$$

Let $b_{d(k,k')}(t)$ be a $L_{(k,k')}$-dimensional binary row-vector that indicates if every column of $\mathcal{K}_{(k,k')}R$ belongs to $\Sigma_{d(k,k')}(t)$, that is, $b_{d(k,k')}(t)(l) = 1$ if $s_l \in \Sigma_{d(k,k')}(t)$, and $b_{d(k,k')}(t)(l) = 0$ otherwise for every $l \in \{1, \ldots, L_{(k,k')}\}$, where s_l is the lth column of $\mathcal{K}_{(k,k')}R$. In this section, we let

$$\mathscr{Z} = \{(z_1, z_2), \ldots, (z_1, z_K), \ldots, (z_{K-1}, z_K)\}.$$

Finally, define $\mathbb{1}(\mathscr{A}) = (1\{(z_1, z_2) \in \mathscr{A}\}, \ldots, 1\{(z_{K-1}, z_K) \in \mathscr{A}\})^T$ for every $\mathscr{A} \subset \mathscr{Z}$.

A.2.2 VSIV Estimation under Consistent Estimation of Validity Pair Set

Here, we study the properties of VSIV Estimation when the validity pair set can be estimated consistently, that is, there is an estimator $\widehat{\mathscr{Z}_0}$ such that $\mathbb{P}(\widehat{\mathscr{Z}_0} = \mathscr{Z}_{\bar{M}}) \to 1$. Suppose that there are subsets $\mathscr{Z}_1 \subset \mathscr{Z}$ and $\mathscr{Z}_2 \subset \mathscr{Z}$ that satisfy the testable implications in Sun (2021), and those in Kédagni and Mourifié (2020), respectively. Similarly to Section A.1, we let $\mathscr{Z}_0 = \mathscr{Z}_1 \cap \mathscr{Z}_2$ so that $\mathscr{Z}_0$ satisfies all the above necessary conditions. We first construct the estimators $\widehat{\mathscr{Z}_1}$ and $\widehat{\mathscr{Z}_2}$ for $\mathscr{Z}_1$ and $\mathscr{Z}_2$, respectively, and then construct the estimator $\widehat{\mathscr{Z}_0}$ for $\mathscr{Z}_0$ as $\widehat{\mathscr{Z}_0} = \widehat{\mathscr{Z}_1} \cap \widehat{\mathscr{Z}_2}$. See Appendix C.2 for details. Under mild conditions, $\mathbb{P}(\widehat{\mathscr{Z}_0} = \mathscr{Z}_0) \to 1$. If $\mathscr{Z}_0 = \mathscr{Z}_{\bar{M}}$, then it follows that $\mathbb{P}(\widehat{\mathscr{Z}_0} = \mathscr{Z}_{\bar{M}}) \to 1$.

To state the results, define

$$P_{DZ}(d) = \left(\mathbb{P}(D = d, Z = z_1), \ldots, \mathbb{P}(D = d, Z = z_K)\right)^T,$$

$$Q_{YDZ}(d) = \left(E[\kappa(Y) 1\{D = d, Z = z_1\}], \ldots, E[\kappa(Y) 1\{D = d, Z = z_K\}]\right)^T,$$

for every $d \in \mathcal{D}$, and

$$Z_P = \left(\mathbb{P}(Z = z_1), \ldots, \mathbb{P}(Z = z_K)\right),$$

$$W = \left(Z_P, P_{DZ}(d_1)^T, \ldots, P_{DZ}(d_J)^T, Q_{YDZ}(d_1)^T, \ldots, Q_{YDZ}(d_J)^T\right)^T.$$

Suppose we have a random sample $\{(Y_i, D_i, Z_i)\}_{i=1}^n$. Define the following sample analogs:

$$\widehat{\mathbb{P}}(Z = z) = \frac{1}{n}\sum_{i=1}^n 1\{Z_i = z\} \text{ for all } z,$$

$$\widehat{\mathbb{P}}(D = d, Z = z) = \frac{1}{n}\sum_{i=1}^n 1\{D_i = d, Z_i = z\} \text{ for all } d \text{ and all } z,$$

$$\widehat{E}[\kappa(Y) 1\{D = d, Z = z\}] = \frac{1}{n}\sum_{i=1}^n \kappa(Y_i) 1\{D_i = d, Z_i = z\} \text{ for all } d \text{ and all } z,$$

$$\widehat{P_{DZ}(d)} = \left(\widehat{\mathbb{P}}(D = d, Z = z_1), \ldots, \widehat{\mathbb{P}}(D = d, Z = z_K)\right)^T \text{ for all } d,$$

$$\widehat{Q_{YDZ}(d)} = \left(\widehat{E}[\kappa(Y) 1\{D = d, Z = z_1\}], \ldots, \widehat{E}[\kappa(Y) 1\{D = d, Z = z_K\}]\right)^T \text{ for all } d,$$

$$\widehat{Z_P} = \left(\widehat{\mathbb{P}}(Z = z_1), \ldots, \widehat{\mathbb{P}}(Z = z_K)\right),$$

and

$$\widehat{W} = \left(\widehat{Z_P}, \widehat{P_{DZ}(d_1)}^T, \ldots, \widehat{P_{DZ}(d_J)}^T, \widehat{Q_{YDZ}(d_1)}^T, \ldots, \widehat{Q_{YDZ}(d_J)}^T\right)^T.$$

We impose the following weak regularity conditions.

Assumption A.6 *$\{(Y_i, D_i, Z_i)\}_{i=1}^n$ is an i.i.d. sample from a population such that all relevant moments exist.*

The next theorem presents the identification and estimation results under pairwise IV validity with unordered treatments.

Theorem A.3 *Suppose that the instrument Z is pairwise valid for the treatment D as defined in Definition A.2 with the largest validity pair set $\mathscr{Z}_{\bar{M}} = \{(z_{k_1}, z_{k'_1}), \ldots, (z_{k_{\bar{M}}}, z_{k'_{\bar{M}}})\}$. Under Assumption A.6, the following response-type probabilities and counterfactuals are identified*

for every $d \in \mathcal{D}$*, each* $t \in \{1,2\}$*, and every* $(z_k, z_{k'}) \in \mathscr{Z}_{\bar{M}}$*:*

$$\mathbb{P}\left(\mathcal{M}_{(k,k')}S \in \Sigma_{d(k,k')}(t)\right) = b_{d(k,k')}(t)\, B^{+}_{d(k,k')} P_{Z(k,k')}(d) \text{ and}$$
$$E[\kappa(Y_d(z_k, z_{k'}))\,|\mathcal{M}_{(k,k')}S \in \Sigma_{d(k,k')}(t)] = \frac{b_{d(k,k')}(t)\, B^{+}_{d(k,k')} Q_{Z(k,k')}(d)}{b_{d(k,k')}(t)\, B^{+}_{d(k,k')} P_{Z(k,k')}(d)}. \tag{A.7}$$

In addition, if $\mathbb{P}(\widehat{\mathscr{Z}_0} = \mathscr{Z}_{\bar{M}}) \to 1$*, we have that*

$$\sqrt{n}\left\{\left(\widehat{W}^T, \mathbb{1}(\widehat{\mathscr{Z}_0})^T\right)^T - \left(W^T, \mathbb{1}(\mathscr{Z}_{\bar{M}})^T\right)^T\right\} \xrightarrow{d} \left(N(0, \Sigma_W)^T, 0^T\right)^T,$$

where Σ_W *is given in* (C.4).

Theorem A.3 is an extension of Theorem T-6 of Heckman and Pinto (2018) for pairwise valid instruments. As shown in Remark 7.1 in Heckman and Pinto (2018) and Theorem A.3, if $(z_k, z_{k'}) \in \mathscr{Z}_{\bar{M}}$ and $\Sigma_{d(k,k')}(t) = \Sigma_{d'(k,k')}(t')$ for some $d, d' \in \mathcal{D}$ and some $t, t' \in \{1,2\}$, the mean treatment effect of d relative to d' for $\Sigma_{d(k,k')}(t)$ can be identified, which is $E[Y_d(z_k, z_{k'}) - Y_{d'}(z_k, z_{k'})|\mathcal{M}_{(k,k')}S \in \Sigma_{d(k,k')}(t)]$.

For all $d, d' \in \mathcal{D}$, all $t, t' \in \{1,2\}$, and all $k < k'$, following Heckman and Pinto (2018), we define

$$\begin{aligned}\beta_{(k,k')}(d, d', t, t') \equiv & 1\{(z_k, z_{k'}) \in \mathscr{Z}_{\bar{M}}, \Sigma_{d(k,k')}(t) = \Sigma_{d'(k,k')}(t')\} \\ & \cdot E[Y_{dz_k} - Y_{d'z_{k'}}|\mathcal{M}_{(k,k')}S \in \Sigma_{d(k,k')}(t)].\end{aligned}$$

When $(z_k, z_{k'}) \in \mathscr{Z}_{\bar{M}}$ and $\Sigma_{d(k,k')}(t) = \Sigma_{d'(k,k')}(t')$, we have that

$$\beta_{(k,k')}(d, d', t, t') = E[Y_d(z_k, z_{k'}) - Y_{d'}(z_k, z_{k'})|\mathcal{M}_{(k,k')}S \in \Sigma_{d(k,k')}(t)],$$

which is the mean treatment effect of d relative to d' for $\Sigma_{d(k,k')}(t)$.

Lemma A.3 *Let* $\kappa(y) = y$ *for all* $y \in \mathbb{R}$*. The mean treatment effect* $\beta_{(k,k')}(d, d', t, t')$ *can be expressed as*

$$\begin{aligned}\beta_{(k,k')}(d, d', t, t') = & 1\{(z_k, z_{k'}) \in \mathscr{Z}_{\bar{M}}, \Sigma_{d(k,k')}(t) = \Sigma_{d'(k,k')}(t')\} \\ & \cdot \left\{\frac{b_{d(k,k')}(t)\, B^{+}_{d(k,k')} Q_{Z(k,k')}(d)}{b_{d(k,k')}(t)\, B^{+}_{d(k,k')} P_{Z(k,k')}(d)} - \frac{b_{d'(k,k')}(t')\, B^{+}_{d'(k,k')} Q_{Z(k,k')}(d')}{b_{d'(k,k')}(t')\, B^{+}_{d'(k,k')} P_{Z(k,k')}(d')}\right\}. \end{aligned} \tag{A.8}$$

We now define

$$\begin{aligned}&\beta_{(k,k')}(d,d')\\&\quad=(\beta_{(k,k')}(d,d',1,1),\beta_{(k,k')}(d,d',1,2),\beta_{(k,k')}(d,d',2,1),\beta_{(k,k')}(d,d',2,2))\end{aligned}\tag{A.9}$$

for all $d,d'\in\mathcal{D}$ and all $k<k'$. For all $k<k'$, we let

$$\beta_{(k,k')}=(\beta_{(k,k')}(d_1,d_2),\ldots,\beta_{(k,k')}(d_1,d_J),\ldots,\beta_{(k,k')}(d_J,d_1),\ldots,\beta_{(k,k')}(d_J,d_{J-1})).$$

Finally, we define

$$\beta=(\beta_{(1,2)},\ldots,\beta_{(1,K)},\ldots,\beta_{(K-1,K)})^T.\tag{A.10}$$

Note that if $(z_k,z_{k'})\notin\mathscr{Z}_{\bar{M}}$, then $\beta_{(k,k')}=0$. For the sample analogs, we define

$$\begin{aligned}\widehat{\beta}_{(k,k')}(d,d',t,t')=&1\{(z_k,z_{k'})\in\widehat{\mathscr{Z}_0},\Sigma_{d(k,k')}(t)=\Sigma_{d'(k,k')}(t')\}\\&\cdot\left\{\frac{b_{d(k,k')}(t)\,B^{+}_{d(k,k')}\widehat{Q_{Z(k,k')}(d)}}{b_{d(k,k')}(t)\,B^{+}_{d(k,k')}\widehat{P_{Z(k,k')}(d)}}-\frac{b_{d'(k,k')}(t')\,B^{+}_{d'(k,k')}\widehat{Q_{Z(k,k')}(d')}}{b_{d'(k,k')}(t')\,B^{+}_{d'(k,k')}\widehat{P_{Z(k,k')}(d')}}\right\},\end{aligned}\tag{A.11}$$

where $\widehat{P_{Z(k,k')}(d)}$ and $\widehat{Q_{Z(k,k')}(d)}$ can be obtained by transformations of $\widehat{W}$. We let

$$\widehat{\beta}_{(k,k')}(d,d')=(\widehat{\beta}_{(k,k')}(d,d',1,1),\widehat{\beta}_{(k,k')}(d,d',1,2),\widehat{\beta}_{(k,k')}(d,d',2,1),\widehat{\beta}_{(k,k')}(d,d',2,2))\tag{A.12}$$

for all $d,d'\in\mathcal{D}$ and all $k<k'$. For all $k<k'$, we define

$$\widehat{\beta}_{(k,k')}=(\widehat{\beta}_{(k,k')}(d_1,d_2),\ldots,\widehat{\beta}_{(k,k')}(d_1,d_K),\ldots,\widehat{\beta}_{(k,k')}(d_K,d_1),\ldots,\widehat{\beta}_{(k,k')}(d_K,d_{K-1})).\tag{A.13}$$

Finally, define

$$\widehat{\beta}=(\widehat{\beta}_{(1,2)},\ldots,\widehat{\beta}_{(1,K)},\ldots,\widehat{\beta}_{(K-1,K)})^T.\tag{A.14}$$

Asymptotic properties of the VSIV estimator in (A.14) can be obtained by Theorem A.3 with $\mathbb{P}(\widehat{\mathscr{Z}_0}=\mathscr{Z}_{\bar{M}})\to 1$.

A.2.3 Bias Reduction for Mean Treatment Effects

Here, we extend the results in Section 3.2 and show that VSIV estimation always reduces the asymptotic bias for estimating mean treatment effects with unordered treatments.

With β and $\widehat{\beta}$ defined in (A.10) and (A.14), the following theorem shows that VSIV estimation always reduces the asymptotic estimation bias.

Theorem A.4 *Suppose that Assumption A.6 holds and that* $\mathbb{P}(\widehat{\mathscr{Z}_0} = \mathscr{Z}_0) \to 1$ *with* $\mathscr{Z}_0 \supset \mathscr{Z}_{\bar{M}}$. *For every presumed validity pair set* $\mathscr{Z}_P$, *the asymptotic bias* $\operatorname{plim}_{n\to\infty}\|\widehat{\beta} - \beta\|_2$ *is always reduced by using* $\widehat{\mathscr{Z}_0'} = \widehat{\mathscr{Z}_0} \cap \mathscr{Z}_P$ *in the estimation for* (A.10) *compared to that from using* $\mathscr{Z}_P$.

As shown in Propositions B.2 and C.1, the pseudo-validity pair set $\mathscr{Z}_0$ can always be estimated consistently by $\widehat{\mathscr{Z}_0}$ under mild conditions. Theorem A.4 shows that VSIV estimation based on $\widehat{\mathscr{Z}_0} \cap \mathscr{Z}_P$ reduces the bias relative to standard IV estimation based on $\mathscr{Z}_P$.

B Proofs and Supplementary Results for Section 2 and Appendix A.1

The results in Section 2 are for the special case where D is binary and follow from the general results for ordered treatments in Appendix A.1. The proofs of these general results are in Appendix B.1.

B.1 Proofs for Appendix A.1

Proof of Lemma A.1. The proof closely follows the strategy of that of Theorem 1 in Angrist and Imbens (1995). Let $d_0 < d_1$ and $Y_{d_0}(z_{k_m}, z_{k'_m}) = 0$ for every m. Let d_{J+1} be some number such that $d_{J+1} > d_J$. We can write

$$Y = \sum_{k=1}^{K} 1\{Z = z_k\} \cdot \left\{ \sum_{j=1}^{J} 1\{D = d_j\} Y_{d_j z_k} \right\}.$$

Now we have that

$$
\begin{aligned}
&E\left[Y|Z=z_{k'_m}\right]-E\left[Y|Z=z_{k_m}\right] \\
=&E\left[\sum_{j=1}^{J}Y_{d_j}\left(z_{k_m},z_{k'_m}\right)\left(\begin{array}{c}1\left\{D_{z_{k'_m}}\geq d_j\right\}-1\left\{D_{z_{k'_m}}\geq d_{j+1}\right\}\\ -1\left\{D_{z_{k_m}}\geq d_j\right\}+1\left\{D_{z_{k_m}}\geq d_{j+1}\right\}\end{array}\right)\right] \\
=&\sum_{j=1}^{J}E\left[\left(Y_{d_j}\left(z_{k_m},z_{k'_m}\right)-Y_{d_{j-1}}\left(z_{k_m},z_{k'_m}\right)\right)\left(1\left\{D_{z_{k'_m}}\geq d_j\right\}-1\left\{D_{z_{k_m}}\geq d_j\right\}\right)\right].
\end{aligned}
$$

By Definition A.1, $(1\{D_{z_{k'_m}}\geq d_j\}-1\{D_{z_{k_m}}\geq d_j\})\in\{0,1\}$. Then we have that

$$
\begin{aligned}
&\sum_{j=1}^{J}E\left[\left(Y_{d_j}\left(z_{k_m},z_{k'_m}\right)-Y_{d_{j-1}}\left(z_{k_m},z_{k'_m}\right)\right)\left(1\left\{D_{z_{k'_m}}\geq d_j\right\}-1\left\{D_{z_{k_m}}\geq d_j\right\}\right)\right] \\
=&\sum_{j=1}^{J}\left\{E\left[\left(Y_{d_j}\left(z_{k_m},z_{k'_m}\right)-Y_{d_{j-1}}\left(z_{k_m},z_{k'_m}\right)\right)|1\left\{D_{z_{k'_m}}\geq d_j\right\}-1\left\{D_{z_{k_m}}\geq d_j\right\}=1\right]\right. \\
&\qquad\left.\cdot\,\mathbb{P}\left(1\left\{D_{z_{k'_m}}\geq d_j\right\}-1\left\{D_{z_{k_m}}\geq d_j\right\}=1\right)\right\} \\
=&\sum_{j=1}^{J}E\left[\left(Y_{d_j}\left(z_{k_m},z_{k'_m}\right)-Y_{d_{j-1}}\left(z_{k_m},z_{k'_m}\right)\right)|D_{z_{k'_m}}\geq d_j>D_{z_{k_m}}\right] \\
&\qquad\cdot\,\mathbb{P}\left(D_{z_{k'_m}}\geq d_j>D_{z_{k_m}}\right).
\end{aligned}
$$

Similarly, we have

$$
\begin{aligned}
&E\left[D|Z=z_{k'_m}\right]-E\left[D|Z=z_{k_m}\right] \\
=&E\left[\sum_{j=1}^{J}d_j\left(1\left\{D_{z_{k'_m}}\geq d_j\right\}-1\left\{D_{z_{k_m}}\geq d_j\right\}\right)\right] \\
&-E\left[\sum_{j=1}^{J}d_j\left(1\left\{D_{z_{k'_m}}\geq d_{j+1}\right\}-1\left\{D_{z_{k_m}}\geq d_{j+1}\right\}\right)\right] \\
=&E\left[\sum_{j=1}^{J}d_j\cdot 1\left\{D_{z_{k'_m}}\geq d_j>D_{z_{k_m}}\right\}\right]-E\left[\sum_{j=1}^{J}d_{j-1}\cdot 1\left\{D_{z_{k'_m}}\geq d_j>D_{z_{k_m}}\right\}\right] \\
=&\sum_{j=1}^{J}\left(d_j-d_{j-1}\right)\mathbb{P}\left(D_{z_{k'_m}}\geq d_j>D_{z_{k_m}}\right).
\end{aligned}
$$

Thus, finally we have that

$$\begin{aligned}\beta_{k'_m,k_m} &\equiv \sum_{j=1}^{J} \omega_j \cdot E\left[\left(Y_{d_j}\left(z_{k_m}, z_{k'_m}\right) - Y_{d_{j-1}}\left(z_{k_m}, z_{k'_m}\right)\right) | D_{z_{k'_m}} \geq d_j > D_{z_{k_m}}\right] \\ &= \frac{E\left[Y|Z = z_{k'_m}\right] - E\left[Y|Z = z_{k_m}\right]}{E\left[D|Z = z_{k'_m}\right] - E\left[D|Z = z_{k_m}\right]},\end{aligned}$$

where

$$\omega_j = \frac{\mathbb{P}\left(D_{z_{k'_m}} \geq d_j > D_{z_{k_m}}\right)}{\sum_{l=1}^{J}\left(d_l - d_{l-1}\right)\mathbb{P}\left(D_{z_{k'_m}} \geq d_l > D_{z_{k_m}}\right)}.$$

Note that by definition, $\mathbb{P}(D_{z_{k'_m}} \geq d_1 > D_{z_{k_m}}) = 0$. ■

Proof of Theorem A.1. For every $\mathcal{Z}_{(k,k')} \in \mathscr{Z}$, we define

$$W_i\left(\mathcal{Z}_{(k,k')}\right) = \begin{pmatrix} g\left(Z_i\right) Y_i 1\left\{Z_i \in \mathcal{Z}_{(k,k')}\right\} \\ Y_i 1\left\{Z_i \in \mathcal{Z}_{(k,k')}\right\} \\ g\left(Z_i\right) 1\left\{Z_i \in \mathcal{Z}_{(k,k')}\right\} \\ g\left(Z_i\right) D_i 1\left\{Z_i \in \mathcal{Z}_{(k,k')}\right\} \\ D_i 1\left\{Z_i \in \mathcal{Z}_{(k,k')}\right\} \\ 1\left\{Z_i \in \mathcal{Z}_{(k,k')}\right\} \end{pmatrix},$$

$$\widehat{W}_n\left(\mathcal{Z}_{(k,k')}\right) = \frac{1}{n}\sum_{i=1}^{n} W_i\left(\mathcal{Z}_{(k,k')}\right), \text{ and } W\left(\mathcal{Z}_{(k,k')}\right) = E\left[W_i\left(\mathcal{Z}_{(k,k')}\right)\right].$$

Also, we let

$$\begin{aligned}&\widehat{W}_n = \left(\widehat{W}_n\left(\mathcal{Z}_{(1,2)}\right)^T, \ldots, \widehat{W}_n\left(\mathcal{Z}_{(1,K)}\right)^T, \ldots, \widehat{W}_n\left(\mathcal{Z}_{(K,1)}\right)^T, \ldots, \widehat{W}_n\left(\mathcal{Z}_{(K,K-1)}\right)^T\right)^T \\ &\text{and } W = \left(W\left(\mathcal{Z}_{(1,2)}\right)^T, \ldots, W\left(\mathcal{Z}_{(1,K)}\right)^T, \ldots, W\left(\mathcal{Z}_{(K,1)}\right)^T, \ldots, W\left(\mathcal{Z}_{(K,K-1)}\right)^T\right)^T.\end{aligned}$$

By multivariate central limit theorem,

$$\begin{aligned}\sqrt{n}\left(\widehat{W}_n - W\right) &= \sqrt{n}\begin{pmatrix} \widehat{W}_n\left(\mathcal{Z}_{(1,2)}\right) - W\left(\mathcal{Z}_{(1,2)}\right) \\ \vdots \\ \widehat{W}_n\left(\mathcal{Z}_{(K,K-1)}\right) - W\left(\mathcal{Z}_{(K,K-1)}\right) \end{pmatrix} \\ &= \sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} W_i\left(\mathcal{Z}_{(1,2)}\right) - W\left(\mathcal{Z}_{(1,2)}\right) \\ \vdots \\ W_i\left(\mathcal{Z}_{(K,K-1)}\right) - W\left(\mathcal{Z}_{(K,K-1)}\right) \end{pmatrix} \xrightarrow{d} N\left(0, \Sigma_P\right), \qquad \text{(B.1)}\end{aligned}$$

where $\Sigma_P = E\left[V_P V_P^T\right]$ and

$$V_P = \begin{pmatrix} W_i\left(\mathcal{Z}_{(1,2)}\right) - W\left(\mathcal{Z}_{(1,2)}\right) \\ \vdots \\ W_i\left(\mathcal{Z}_{(K,K-1)}\right) - W\left(\mathcal{Z}_{(K,K-1)}\right) \end{pmatrix}.$$

Define a function $f : \mathbb{R}^6 \to \mathbb{R} \cup \{\infty\}$ by

$$f(x) = \frac{x_1/x_6 - x_2 x_3/x_6^2}{x_4/x_6 - x_5 x_3/x_6^2}$$

for every $x \in \mathbb{R}^6$ with $x = (x_1, x_2, x_3, x_4, x_5, x_6)^T$ such that $f(x)$ is well defined. We can obtain the gradient of f, denoted f', by $f'(x) = (f_1'(x), f_2'(x), f_3'(x), f_4'(x), f_5'(x), f_6'(x))^T$ with

$$f_1'(x) = \frac{x_6}{x_4 x_6 - x_5 x_3}, f_2'(x) = \frac{-x_3}{x_4 x_6 - x_5 x_3}, f_3'(x) = \frac{-x_2 x_4 x_6 + x_5 x_1 x_6}{(x_4 x_6 - x_5 x_3)^2},$$
$$f_4'(x) = -\frac{(x_1 x_6 - x_2 x_3) x_6}{(x_4 x_6 - x_5 x_3)^2}, f_5'(x) = \frac{x_3 (x_1 x_6 - x_2 x_3)}{(x_4 x_6 - x_5 x_3)^2}, \text{ and } f_6'(x) = \frac{-x_1 x_5 x_3 + x_2 x_3 x_4}{(x_4 x_6 - x_5 x_3)^2}$$

for every $x = (x_1, x_2, x_3, x_4, x_5, x_6)^T$ such that all the above derivatives are well defined.

For every $\mathcal{Z}_{(k,k')}$, by assumption we have that for every $\rho \geq 0$,

$$\mathbb{P}\left(n^\rho \left| 1\left\{\mathcal{Z}_{(k,k')} \in \widehat{\mathscr{Z}_0}\right\} - 1\left\{\mathcal{Z}_{(k,k')} \in \mathscr{Z}_{\bar{M}}\right\}\right| > \varepsilon\right) \leq \mathbb{P}\left(\widehat{\mathscr{Z}_0} \neq \mathscr{Z}_{\bar{M}}\right) \to 0. \tag{B.2}$$

This implies that if $1\{\mathcal{Z}_{(k,k')} \in \mathscr{Z}_{\bar{M}}\} = 0$, then

$$n^\rho 1\{\mathcal{Z}_{(k,k')} \in \widehat{\mathscr{Z}_0}\} = o_p(1). \tag{B.3}$$

Without loss of generality, we suppose $\mathscr{Z}_{\bar{M}} = \{\mathcal{Z}_{(1,2)}, \mathcal{Z}_{(1,3)}, \ldots, \mathcal{Z}_{(K-1,K)}\}$ and $\mathscr{Z} \setminus \mathscr{Z}_{\bar{M}} = \{\mathcal{Z}_{(2,1)}, \mathcal{Z}_{(3,1)}, \ldots, \mathcal{Z}_{(K,K-1)}\}$ for simplicity. For every $\mathcal{Z}_{(k,k')} \notin \mathscr{Z}_{\bar{M}}$, by Assumption A.3, it is possible that

$$E\left[g(Z_i) D_i | Z_i \in \mathcal{Z}_{(k,k')}\right] - E\left[g(Z_i) | Z_i \in \mathcal{Z}_{(k,k')}\right] E\left[D_i | Z_i \in \mathcal{Z}_{(k,k')}\right] = 0. \tag{B.4}$$

For every $w = (w_1^T, \ldots, w_{(K-1)K}^T)^T$ with $w_j = (w_{j1}, \ldots, w_{j6})^T$ for every j, define

$$\begin{aligned}\mathcal{F}_1(w) &= \left(f(w_1), \ldots, f\left(w_{(K-1)K/2}\right)\right)^T \text{ and} \\ \mathcal{F}_0(w) &= \left(f\left(w_{K(K-1)/2+1}\right), \ldots, f\left(w_{(K-1)K}\right)\right)^T.\end{aligned}$$

For every $\mathscr{Z}_s \subset \mathscr{Z}$, define

$$\mathcal{I}_1(\mathscr{Z}_s) = \begin{pmatrix} 1\{\mathcal{Z}_{(1,2)} \in \mathscr{Z}_s\} & & & \\ & 1\{\mathcal{Z}_{(1,3)} \in \mathscr{Z}_s\} & & \\ & & \ddots & \\ & & & 1\{\mathcal{Z}_{(K-1,K)} \in \mathscr{Z}_s\} \end{pmatrix}$$

and

$$\mathcal{I}_0(\mathscr{Z}_s) = \begin{pmatrix} 1\{\mathcal{Z}_{(2,1)} \in \mathscr{Z}_s\} & & & \\ & 1\{\mathcal{Z}_{(3,1)} \in \mathscr{Z}_s\} & & \\ & & \ddots & \\ & & & 1\{\mathcal{Z}_{(K,K-1)} \in \mathscr{Z}_s\} \end{pmatrix}.$$

Then we can write

$$\sqrt{n}\left(\widehat{\beta}_1 - \beta_1\right) = \sqrt{n}\left\{\begin{pmatrix} \mathcal{I}_1\left(\widehat{\mathscr{Z}_0}\right)\mathcal{F}_1\left(\widehat{W}_n\right) \\ \mathcal{I}_0\left(\widehat{\mathscr{Z}_0}\right)\mathcal{F}_0\left(\widehat{W}_n\right) \end{pmatrix} - \begin{pmatrix} \mathcal{I}_1(\mathscr{Z}_{\bar{M}})\mathcal{F}_1(W) \\ \mathcal{I}_0(\mathscr{Z}_{\bar{M}})\mathcal{F}_0(W) \end{pmatrix}\right\}.$$

First, we have that

$$\begin{aligned}\sqrt{n}\left\{\mathcal{I}_1\left(\widehat{\mathscr{Z}_0}\right)\mathcal{F}_1\left(\widehat{W}_n\right) - \mathcal{I}_1(\mathscr{Z}_{\bar{M}})\mathcal{F}_1(W)\right\} &= \sqrt{n}\left\{\mathcal{I}_1\left(\widehat{\mathscr{Z}_0}\right)\mathcal{F}_1\left(\widehat{W}_n\right) - \mathcal{I}_1\left(\widehat{\mathscr{Z}_0}\right)\mathcal{F}_1(W)\right\} \\ &\quad + \sqrt{n}\left\{\mathcal{I}_1\left(\widehat{\mathscr{Z}_0}\right)\mathcal{F}_1(W) - \mathcal{I}_1(\mathscr{Z}_{\bar{M}})\mathcal{F}_1(W)\right\}.\end{aligned}$$

The Jacobian matrix $\mathcal{F}_1'(W)$ of $\mathcal{F}_1$ at W can be obtained with the derivatives of f. Then by (B.2) and delta method, it is easy to show that

$$\begin{aligned}\sqrt{n}\left\{\mathcal{I}_1\left(\widehat{\mathscr{Z}_0}\right)\mathcal{F}_1\left(\widehat{W}_n\right) - \mathcal{I}_1(\mathscr{Z}_{\bar{M}})\mathcal{F}_1(W)\right\} &= \mathcal{I}_1\left(\widehat{\mathscr{Z}_0}\right)\sqrt{n}\left\{\mathcal{F}_1\left(\widehat{W}_n\right) - \mathcal{F}_1(W)\right\} + o_p(1) \\ &\xrightarrow{d} \mathcal{I}_1(\mathscr{Z}_{\bar{M}})\mathcal{F}_1'(W)N(0, \Sigma_P).\end{aligned}$$

Second, by assumption and (1.1),

$$\sqrt{n}\left\{\mathcal{I}_0\left(\widehat{\mathscr{Z}_0}\right)\mathcal{F}_0\left(\widehat{W}_n\right)-\mathcal{I}_0\left(\mathscr{Z}_{\bar{M}}\right)\mathcal{F}_0\left(W\right)\right\}=\sqrt{n}\mathcal{I}_0\left(\widehat{\mathscr{Z}_0}\right)\mathcal{F}_0\left(\widehat{W}_n\right).$$

For every $\mathcal{Z}_{(k,k')} \notin \mathscr{Z}_{\bar{M}}$ such that (B.4) holds,

$$\sqrt{n}1\{\mathcal{Z}_{(k,k')} \in \widehat{\mathscr{Z}_0}\}f(\widehat{W}_n(\mathcal{Z}_{(k,k')})) = n1\{\mathcal{Z}_{(k,k')} \in \widehat{\mathscr{Z}_0}\}\frac{A_n}{\sqrt{n}B_n},$$

where

$$\begin{aligned}A_n =& \frac{1}{n}\sum_{i=1}^{n} g\left(Z_i\right)Y_i 1\left\{Z_i \in \mathcal{Z}_{(k,k')}\right\}\frac{1}{n}\sum_{i=1}^{n} 1\left\{Z_i \in \mathcal{Z}_{(k,k')}\right\}\\ &-\frac{1}{n}\sum_{i=1}^{n} g\left(Z_i\right)1\left\{Z_i \in \mathcal{Z}_{(k,k')}\right\}\frac{1}{n}\sum_{i=1}^{n} Y_i 1\left\{Z_i \in \mathcal{Z}_{(k,k')}\right\}\end{aligned}$$

and

$$\begin{aligned}B_n =& \frac{1}{n}\sum_{i=1}^{n} g\left(Z_i\right)D_i 1\left\{Z_i \in \mathcal{Z}_{(k,k')}\right\}\frac{1}{n}\sum_{i=1}^{n} 1\left\{Z_i \in \mathcal{Z}_{(k,k')}\right\}\\ &-\frac{1}{n}\sum_{i=1}^{n} g\left(Z_i\right)1\left\{Z_i \in \mathcal{Z}_{(k,k')}\right\}\frac{1}{n}\sum_{i=1}^{n} D_i 1\left\{Z_i \in \mathcal{Z}_{(k,k')}\right\}.\end{aligned}$$

Define a map h such that for every $x \in \mathbb{R}^6$ with $x = (x_1, \ldots, x_6)^T$,

$$h\left(x\right) = x_4x_6 - x_3x_5.$$

Let $h'(W(\mathcal{Z}_{(k,k')}))$ be the Jacobian matrix of h at $W(\mathcal{Z}_{(k,k')})$. Then by delta method,

$$\sqrt{n}B_n = \sqrt{n}\left(h\left(\widehat{W}_n\left(\mathcal{Z}_{(k,k')}\right)\right)-h\left(W\left(\mathcal{Z}_{(k,k')}\right)\right)\right)\xrightarrow{d} h'\left(W\left(\mathcal{Z}_{(k,k')}\right)\right)N\left(0,\Sigma_{(k,k')}\right),$$

where

$$\Sigma_{(k,k')} = E\left[\left\{W_i\left(\mathcal{Z}_{(k,k')}\right)-W\left(\mathcal{Z}_{(k,k')}\right)\right\}\left\{W_i\left(\mathcal{Z}_{(k,k')}\right)-W\left(\mathcal{Z}_{(k,k')}\right)\right\}^T\right].$$

Also, it is easy to show that

$$\begin{aligned}A_n \xrightarrow{p}& E\left[g\left(Z_i\right)Y_i 1\left\{Z_i \in \mathcal{Z}_{(k,k')}\right\}\right]E\left[1\left\{Z_i \in \mathcal{Z}_{(k,k')}\right\}\right]\\ &-E\left[g\left(Z_i\right)1\left\{Z_i \in \mathcal{Z}_{(k,k')}\right\}\right]E\left[Y_i 1\left\{Z_i \in \mathcal{Z}_{(k,k')}\right\}\right].\end{aligned}$$

Notice that by (B.3), $n\mathcal{I}_0(\widehat{\mathscr{Z}_0}) = o_p(1)$. Thus, $\sqrt{n}1\{\mathcal{Z}_{(k,k')} \in \widehat{\mathscr{Z}_0}\} f(\widehat{W}_n(\mathcal{Z}_{(k,k')})) \xrightarrow{p} 0$. Similarly, for every $\mathcal{Z}_{(k,k')} \notin \mathscr{Z}_{\bar{M}}$ such that (B.4) does not hold, it is easy to show that

$$\sqrt{n}1\{\mathcal{Z}_{(k,k')} \in \widehat{\mathscr{Z}_0}\} f(\widehat{W}_n(\mathcal{Z}_{(k,k')})) = \sqrt{n}1\{\mathcal{Z}_{(k,k')} \in \widehat{\mathscr{Z}_0}\} \frac{A_n}{B_n} \xrightarrow{p} 0.$$

This implies that

$$\sqrt{n}\left\{\mathcal{I}_0\left(\widehat{\mathscr{Z}_0}\right)\mathcal{F}_0\left(\widehat{W}_n\right) - \mathcal{I}_0\left(\mathscr{Z}_{\bar{M}}\right)\mathcal{F}_0\left(W\right)\right\} \xrightarrow{p} 0.$$

By Lemma 1.10.2(iii) and Example 1.4.7 (Slutsky's lemma) of van der Vaart and Wellner (1996),

$$\begin{aligned}\sqrt{n}\left(\widehat{\beta}_1 - \beta_1\right) &= \sqrt{n}\left\{\begin{pmatrix} \mathcal{I}_1\left(\widehat{\mathscr{Z}_0}\right)\mathcal{F}_1\left(\widehat{W}_n\right) \\ \mathcal{I}_0\left(\widehat{\mathscr{Z}_0}\right)\mathcal{F}_0\left(\widehat{W}_n\right)\end{pmatrix} - \begin{pmatrix} \mathcal{I}_1\left(\mathscr{Z}_{\bar{M}}\right)\mathcal{F}_1\left(W\right) \\ \mathcal{I}_0\left(\mathscr{Z}_{\bar{M}}\right)\mathcal{F}_0\left(W\right)\end{pmatrix}\right\} \\ &\xrightarrow{d} \begin{pmatrix} \mathcal{I}_1\left(\mathscr{Z}_{\bar{M}}\right)\mathcal{F}_1'\left(W\right)N\left(0,\Sigma_P\right) \\ 0\end{pmatrix}. \end{aligned} \tag{B.5}$$

Now we have that for every $\mathcal{Z}_{(k,k')} \in \mathscr{Z}_{\bar{M}}$,

$$\begin{aligned}&\frac{E\left[g\left(Z_i\right)Y_i 1\left\{Z_i \in \mathcal{Z}_{(k,k')}\right\}\right]}{\mathbb{P}\left(Z_i \in \mathcal{Z}_{(k,k')}\right)} - \frac{E\left[Y_i 1\left\{Z_i \in \mathcal{Z}_{(k,k')}\right\}\right]}{\mathbb{P}\left(Z_i \in \mathcal{Z}_{(k,k')}\right)}\frac{E\left[g\left(Z_i\right)1\left\{Z_i \in \mathcal{Z}_{(k,k')}\right\}\right]}{\mathbb{P}\left(Z_i \in \mathcal{Z}_{(k,k')}\right)} \\ =&\sum_{l=1}^{K}\left\{\begin{array}{c} \frac{\mathbb{P}(Z_i=z_l)}{\mathbb{P}\left(Z_i \in \mathcal{Z}_{(k,k')}\right)} E\left[Y_i 1\left\{Z_i \in \mathcal{Z}_{(k,k')}\right\}|Z_i = z_l\right] \\ \cdot\left\{g\left(z_l\right)1\left\{z_l \in \mathcal{Z}_{(k,k')}\right\} - \frac{E\left[g(Z_i)1\left\{Z_i \in \mathcal{Z}_{(k,k')}\right\}\right]}{\mathbb{P}\left(Z_i \in \mathcal{Z}_{(k,k')}\right)}\right\}\end{array}\right\} \\ =&\mathbb{P}\left(Z_i = z_k | Z_i \in \mathcal{Z}_{(k,k')}\right) E\left[Y_i|Z_i = z_k\right]\left\{g\left(z_k\right) - E\left[g\left(Z_i\right)|Z_i \in \mathcal{Z}_{(k,k')}\right]\right\} \\ &+\mathbb{P}\left(Z_i = z_{k'} | Z_i \in \mathcal{Z}_{(k,k')}\right) E\left[Y_i|Z_i = z_{k'}\right]\left\{g\left(z_{k'}\right) - E\left[g\left(Z_i\right)|Z_i \in \mathcal{Z}_{(k,k')}\right]\right\}.\end{aligned}$$

By (A.1), we have

$$E\left[Y_i|Z_i = z_{k'}\right] = \beta_{k',k}\left(E\left[D_i|Z_i = z_{k'}\right] - E\left[D_i|Z_i = z_k\right]\right) + E\left[Y_i|Z_i = z_k\right],$$

and thus it follows that

$$\begin{aligned}
&\mathbb{P}\left(Z_i = z_k | Z_i \in \mathcal{Z}_{(k,k')}\right) E\left[Y_i | Z_i = z_k\right] \left\{g\left(z_k\right) - E\left[g\left(Z_i\right) | Z_i \in \mathcal{Z}_{(k,k')}\right]\right\} \\
&+ \mathbb{P}\left(Z_i = z_{k'} | Z_i \in \mathcal{Z}_{(k,k')}\right) E\left[Y_i | Z_i = z_{k'}\right] \left\{g\left(z_{k'}\right) - E\left[g\left(Z_i\right) | Z_i \in \mathcal{Z}_{(k,k')}\right]\right\} \\
=&\mathbb{P}\left(Z_i = z_{k'} | Z_i \in \mathcal{Z}_{(k,k')}\right) \beta_{k',k} \left(E\left[D_i | Z_i = z_{k'}\right] - E\left[D_i | Z_i = z_k\right]\right) \\
&\cdot \left\{g\left(z_{k'}\right) - E\left[g\left(Z_i\right) | Z_i \in \mathcal{Z}_{(k,k')}\right]\right\},
\end{aligned}$$

where we use the equality that

$$\begin{aligned}
&\mathbb{P}\left(Z_i = z_k | Z_i \in \mathcal{Z}_{(k,k')}\right) \left\{g\left(z_k\right) - E\left[g\left(Z_i\right) | Z_i \in \mathcal{Z}_{(k,k')}\right]\right\} \\
&+ \mathbb{P}\left(Z_i = z_{k'} | Z_i \in \mathcal{Z}_{(k,k')}\right) \left\{g\left(z_{k'}\right) - E\left[g\left(Z_i\right) | Z_i \in \mathcal{Z}_{(k,k')}\right]\right\} = 0. \qquad \text{(B.6)}
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
&\frac{E\left[g\left(Z_i\right) D_i 1\left\{Z_i \in \mathcal{Z}_{(k,k')}\right\}\right]}{\mathbb{P}\left(Z_i \in \mathcal{Z}_{(k,k')}\right)} - \frac{E\left[D_i 1\left\{Z_i \in \mathcal{Z}_{(k,k')}\right\}\right]}{\mathbb{P}\left(Z_i \in \mathcal{Z}_{(k,k')}\right)} \frac{E\left[g\left(Z_i\right) 1\left\{Z_i \in \mathcal{Z}_{(k,k')}\right\}\right]}{\mathbb{P}\left(Z_i \in \mathcal{Z}_{(k,k')}\right)} \\
=&\mathbb{P}\left(Z_i = z_{k'} | Z_i \in \mathcal{Z}_{(k,k')}\right) \left\{p(z_{k'}) - p(z_k)\right\} \left\{g\left(z_{k'}\right) - E\left[g\left(Z_i\right) | Z_i \in \mathcal{Z}_{(k,k')}\right]\right\},
\end{aligned}$$

where $p(z) = E\left[D_i | Z_i = z\right]$ for all z and we use the equality in (B.6) again. ■

Proof of Theorem A.2. Recall that for every random variable ξ_i and every $\mathcal{A} \in \mathscr{Z}$,

$$\mathcal{E}_n\left(\xi_i, \mathcal{A}\right) = \frac{\frac{1}{n}\sum_{i=1}^n \xi_i 1\left\{Z_i \in \mathcal{A}\right\}}{\frac{1}{n}\sum_{i=1}^n 1\left\{Z_i \in \mathcal{A}\right\}} \text{ and } \mathcal{E}\left(\xi_i, \mathcal{A}\right) = \frac{E\left[\xi_i 1\left\{Z_i \in \mathcal{A}\right\}\right]}{E\left[1\left\{Z_i \in \mathcal{A}\right\}\right]}.$$

Then we obtain the VSIV estimator using $\mathscr{Z}_P$ for each ACR as

$$\widehat{\beta}'_{(k,k')} = 1\{\mathcal{Z}_{(k,k')} \in \mathscr{Z}_P\} \cdot \frac{\mathcal{E}_n\left(g\left(Z_i\right) Y_i, \mathcal{Z}_{(k,k')}\right) - \mathcal{E}_n\left(g\left(Z_i\right), \mathcal{Z}_{(k,k')}\right) \mathcal{E}_n\left(Y_i, \mathcal{Z}_{(k,k')}\right)}{\mathcal{E}_n\left(g\left(Z_i\right) D_i, \mathcal{Z}_{(k,k')}\right) - \mathcal{E}_n\left(g\left(Z_i\right), \mathcal{Z}_{(k,k')}\right) \mathcal{E}_n\left(D_i, \mathcal{Z}_{(k,k')}\right)},$$

which converges in probability to

$$\beta'_{(k,k')} = 1\{\mathcal{Z}_{(k,k')} \in \mathscr{Z}_P\} \cdot \frac{\mathcal{E}\left(g\left(Z_i\right) Y_i, \mathcal{Z}_{(k,k')}\right) - \mathcal{E}\left(g\left(Z_i\right), \mathcal{Z}_{(k,k')}\right) \mathcal{E}\left(Y_i, \mathcal{Z}_{(k,k')}\right)}{\mathcal{E}\left(g\left(Z_i\right) D_i, \mathcal{Z}_{(k,k')}\right) - \mathcal{E}\left(g\left(Z_i\right), \mathcal{Z}_{(k,k')}\right) \mathcal{E}\left(D_i, \mathcal{Z}_{(k,k')}\right)}.$$

We obtain the VSIV estimator using $\widehat{\mathscr{Z}}'_0$ for each ACR as

$$\widehat{\beta}''_{(k,k')} = 1\{\mathcal{Z}_{(k,k')} \in \widehat{\mathscr{Z}}'_0\} \cdot \frac{\mathcal{E}_n\left(g\left(Z_i\right) Y_i, \mathcal{Z}_{(k,k')}\right) - \mathcal{E}_n\left(g\left(Z_i\right), \mathcal{Z}_{(k,k')}\right) \mathcal{E}_n\left(Y_i, \mathcal{Z}_{(k,k')}\right)}{\mathcal{E}_n\left(g\left(Z_i\right) D_i, \mathcal{Z}_{(k,k')}\right) - \mathcal{E}_n\left(g\left(Z_i\right), \mathcal{Z}_{(k,k')}\right) \mathcal{E}_n\left(D_i, \mathcal{Z}_{(k,k')}\right)},$$

which converges in probability to

$$\beta''_{(k,k')} = 1\{\mathcal{Z}_{(k,k')} \in \mathscr{Z}'_0\} \cdot \frac{\mathcal{E}\left(g\left(Z_i\right) Y_i, \mathcal{Z}_{(k,k')}\right) - \mathcal{E}\left(g\left(Z_i\right), \mathcal{Z}_{(k,k')}\right) \mathcal{E}\left(Y_i, \mathcal{Z}_{(k,k')}\right)}{\mathcal{E}\left(g\left(Z_i\right) D_i, \mathcal{Z}_{(k,k')}\right) - \mathcal{E}\left(g\left(Z_i\right), \mathcal{Z}_{(k,k')}\right) \mathcal{E}\left(D_i, \mathcal{Z}_{(k,k')}\right)},$$

where $\mathscr{Z}'_0 = \mathscr{Z}_0 \cap \mathscr{Z}_P$.

If $\mathcal{Z}_{(k,k')} \notin \mathscr{Z}_{\bar{M}}$ and $\mathcal{Z}_{(k,k')} \in \mathscr{Z}_P$, then $\beta^1_{(k,k')} = 0$. In this case, it is possible that $\mathcal{Z}_{(k,k')} \notin \mathscr{Z}'_0$ and $\beta''_{(k,k')} = 0$, because by definition $\mathscr{Z}'_0 \subset \mathscr{Z}_P$. Note that if $\mathcal{Z}_{(k,k')} \in \mathscr{Z}'_0$, then $\beta''_{(k,k')} = \beta'_{(k,k')}$ by definition.

If $\mathcal{Z}_{(k,k')} \notin \mathscr{Z}_{\bar{M}}$ and $\mathcal{Z}_{(k,k')} \notin \mathscr{Z}_P$, then $\beta^1_{(k,k')} = \beta'_{(k,k')} = 0$. Similarly, in this case, $\beta''_{(k,k')} = \beta^1_{(k,k')} = 0$, because $\mathscr{Z}'_0 \subset \mathscr{Z}_P$.

If $\mathcal{Z}_{(k,k')} \in \mathscr{Z}_{\bar{M}}$ and $\mathcal{Z}_{(k,k')} \in \mathscr{Z}_P$, then $\beta^1_{(k,k')} = \beta'_{(k,k')} = \beta''_{(k,k')}$, because $\mathscr{Z}_0 \supset \mathscr{Z}_{\bar{M}}$.

If $\mathcal{Z}_{(k,k')} \in \mathscr{Z}_{\bar{M}}$ and $\mathcal{Z}_{(k,k')} \notin \mathscr{Z}_P$, then $\beta'_{(k,k')} = \beta''_{(k,k')} = 0$ because $\mathscr{Z}'_0 \subset \mathscr{Z}_P$. ■

Proposition 3.1 can straightforwardly be extended to multivalued ordered D. We omit this extension here.

Proof of Proposition 3.1. If H_0 is true, it can be shown that under the assumptions,

$$\mathbb{P}\left(\{TS_{1n} = 0\} \cup \{TS_{2n} > c_r(\alpha)\}\right) \geq \mathbb{P}\left(TS_{2n} > c_r(\alpha)\right) \to \alpha$$

and

$$\begin{aligned}\mathbb{P}\left(\{TS_{1n} = 0\} \cup \{TS_{2n} > c_r(\alpha)\}\right) &\leq \mathbb{P}\left(TS_{2n} > c_r(\alpha)\right) + \mathbb{P}\left(TS_{1n} = 0\right) \\ &\leq \mathbb{P}\left(TS_{2n} > c_r(\alpha)\right) + \mathbb{P}(\widehat{\mathscr{Z}_0} \neq \mathscr{Z}_{\bar{M}}) \to \alpha,\end{aligned}$$

which imply that $\mathbb{P}\left(\{TS_{1n} = 0\} \cup \{TS_{2n} > c_r(\alpha)\}\right) \to \alpha$.

Suppose H_0 is false. If $\mathcal{Z}_{(\kappa_m,\kappa'_m)} \notin \mathscr{Z}_{\bar{M}}$ for some m, then

$$\mathbb{P}\left(\{TS_{1n} = 0\} \cup \{TS_{2n} > c_r(\alpha)\}\right) \geq \mathbb{P}\left(TS_{1n} = 0\right) \geq \mathbb{P}(\widehat{\mathscr{Z}_0} = \mathscr{Z}_{\bar{M}}) \to 1.$$

If $\mathcal{Z}_{(\kappa_m,\kappa'_m)} \in \mathscr{Z}_{\bar{M}}$ for all $m \in \{1, \ldots, S\}$ but $R(\beta_{1S}) \neq 0$, then

$$\mathbb{P}\left(\{TS_{1n} = 0\} \cup \{TS_{2n} > c_r(\alpha)\}\right) \geq \mathbb{P}\left(TS_{2n} > c_r(\alpha)\right) \to 1.$$

■

B.2 Selectively Pairwise Valid Multiple Instruments

Here we introduce a weaker notion of pairwise validity that is available when Z contains multiple instruments. Specifically, suppose the instrument Z is a vector with $Z = (Z_1, \ldots, Z_L)^T$, where Z_l is a scalar instrument for every $l \in \{1, \ldots, L\}$. There are $C_L = 2^L$ combinations of scalar instruments $\{Z_1, \ldots, Z_L\}$. We refer to each combination as a *subinstrument* of Z, denoted by V_c for every $c \in \{1, \ldots, C_L\}$ with $V_c \in \{v_{c1}, \ldots, v_{cK_c}\}$ for some $K_c > 1$. Every V_c can be a scalar or vector instrument, and we define the set of all pairs of values of V_c by

$$\mathscr{Z}_c = \{(v_{c1}, v_{c2}), \ldots, (v_{c1}, v_{cK_c}), \ldots, (v_{cK_c}, v_{c1}), \ldots, (v_{cK_c}, v_{cK_c-1})\}.$$

The following definition weakens Definition 2.1.

Definition B.1 *The instrument Z is **selectively pairwise valid** for the treatment $D \in \{0, 1\}$ if there is a subinstrument V_c that is pairwise valid according to Definition 2.1.*

To illustrate that Definition B.1 is weaker than Definition 2.1, consider the following example.

Example B.1 *Suppose that $Z = (Z_1, Z_2, Z_3)^T$, where Z_1 is correlated with all potential variables and $(Z_2, Z_3)^T$ satisfies the conditions in Assumption 2.1. Then Z may not be pairwise valid by Definition 2.1, but it is selectively pairwise valid.*

For every subinstrument V_c, we can define the largest validity pair set $\mathscr{Z}_{c\bar{M}_c} \subset \mathscr{Z}_c$. Then the identification and estimation of $\mathscr{Z}_{c\bar{M}_c}$ and the VSIV estimation of the treatment effects can proceed as described in Section 3.1. Asymptotic normality and bias reduction can be established accordingly. The notion of selectively pairwise valid instruments can be straightforwardly generalized to multivalued ordered or unordered treatments.

B.3 Testable Implications of Kédagni and Mourifié (2020)

We consider the case where $D \in \mathcal{D} = \{d_1, \ldots, d_J\}$. Suppose $Y \in \mathbb{R}$ is continuous. Results for discrete Y can be obtained similarly. The testable implications in Kédagni and Mourifié (2020) are for exclusion ($Y_{dz_{k_m}} = Y_{dz_{k'_m}}$ for all $d \in \mathcal{D}$) and statistical independence ($(Y_{d_1 z_{k_m}}, Y_{d_1 z_{k'_m}}, \ldots, Y_{d_J z_{k_m}}, Y_{d_J z_{k'_m}}) \perp Z$) for every $m \in \{1, \ldots, \bar{M}\}$ with the largest validity pair set $\mathscr{Z}_{\bar{M}} = \{(z_{k_1}, z_{k'_1}), \ldots, (z_{k_{\bar{M}}}, z_{k'_{\bar{M}}})\}$. In the following, we show that these testable

implications are also for Conditions (i) and (ii) in Definition A.1. Under Conditions (i) and (ii) in Definition A.1, we can define $Y_d(z, z')$ by $Y_d(z, z') = Y_{dz} = Y_{dz'}$ a.s. for every $d \in \mathcal{D}$ and every $(z, z') \in \mathscr{Z}_{\bar{M}}$. Define

$$f_{Y,D}(y, d|z) = f_{Y|D,Z}(y|d, z)\,\mathbb{P}(D = d|Z = z)$$

for every $y \in \mathbb{R}$, every $d \in \mathcal{D}$, and every $z \in \mathcal{Z}$, where $f_{Y|D,Z}(y|d, z)$ is the conditional density function of Y given $D = d$ and $Z = z$. For every $\mathcal{Z}_{(k,k')} = (z_k, z_{k'}) \in \mathscr{Z}_{\bar{M}}$, every $A \in \mathcal{B}_{\mathbb{R}}$, every $d \in \mathcal{D}$, and each $z \in \mathcal{Z}_{(k,k')}$,

$$\mathbb{P}(Y \in A, D = d|Z = z) \leq \mathbb{P}(Y_{dz} \in A|Z = z) = \mathbb{P}(Y_d(z_k, z_{k'}) \in A),$$

and

$$\begin{aligned}\mathbb{P}(Y \in A, D = d|Z = z) &= \frac{\mathbb{P}(Y \in A, D = d, Z = z)}{\mathbb{P}(Z = z)} \\ &= \mathbb{P}(Y \in A|D = d, Z = z)\,\mathbb{P}(D = d|Z = z).\end{aligned}$$

Then, by the discussion in Section 4.1 of Kédagni and Mourifié (2020), for (almost) all y,

$$f_{Y,D}(y, d|z) = f_{Y|D,Z}(y|d, z)\,\mathbb{P}(D = d|Z = z) \leq f_{Y_d(z_k,z_{k'})}(y),$$

where $f_{Y_d(z_k,z_{k'})}$ is the density function of the potential outcome $Y_d(z_k, z_{k'})$. Thus, for every $d \in \mathcal{D}$,

$$\max_{z \in \mathcal{Z}_{(k,k')}} f_{Y,D}(y, d|z) \leq f_{Y_d(z_k,z_{k'})}(y), \tag{B.7}$$

and we obtain the first inequality of Kédagni and Mourifié (2020):

$$\max_{d \in \mathcal{D}} \int_{\mathbb{R}} \max_{z \in \mathcal{Z}_{(k,k')}} f_{Y,D}(y, d|z)\,\mathrm{d}y \leq 1. \tag{B.8}$$

Also, for all $A_1, \ldots, A_J \in \mathcal{B}_{\mathbb{R}}$,

$$\begin{aligned}
&\mathbb{P}\left(Y_{d_1}(z_k, z_{k'}) \in A_1, \ldots, Y_{d_J}(z_k, z_{k'}) \in A_J\right) \\
&= \min_{z \in \mathcal{Z}_{(k,k')}} \mathbb{P}\left(Y_{d_1}(z_k, z_{k'}) \in A_1, \ldots, Y_{d_J}(z_k, z_{k'}) \in A_J | Z = z\right) \\
&= \min_{z \in \mathcal{Z}_{(k,k')}} \sum_{j=1}^{J} \mathbb{P}\left(Y_{d_1}(z_k, z_{k'}) \in A_1, \ldots, Y_{d_J}(z_k, z_{k'}) \in A_J, D = d_j | Z = z\right) \\
&\leq \min_{z \in \mathcal{Z}_{(k,k')}} \sum_{j=1}^{J} \mathbb{P}\left(Y \in A_j, D = d_j | Z = z\right).
\end{aligned}$$

Let $P_{\mathbb{R}}^j$ be an arbitrary partition of $\mathbb{R}$ for $j \in \{1, \ldots, J\}$, that is, $P_{\mathbb{R}}^j = \{C_1^j, \ldots, C_{N_j}^j\}$ with $\cup_{l=1}^{N_j} C_l^j = \mathbb{R}$ and $C_{l'}^j \cap C_l^j = \varnothing$ for all $l' \neq l$. Then

$$\begin{aligned}
1 &= \sum_{A_1 \in P_{\mathbb{R}}^1} \cdots \sum_{A_J \in P_{\mathbb{R}}^J} \mathbb{P}\left(Y_{d_1}(z_k, z_{k'}) \in A_1, \ldots, Y_{d_J}(z_k, z_{k'}) \in A_J\right) \\
&\leq \sum_{A_1 \in P_{\mathbb{R}}^1} \cdots \sum_{A_J \in P_{\mathbb{R}}^J} \min_{z \in \mathcal{Z}_{(k,k')}} \sum_{j=1}^{J} \mathbb{P}\left(Y \in A_j, D = d_j | Z = z\right).
\end{aligned}$$

Then we obtain the second inequality of Kédagni and Mourifié (2020):

$$\inf_{\{P_{\mathbb{R}}^1, \ldots, P_{\mathbb{R}}^J\}} \sum_{A_1 \in P_{\mathbb{R}}^1} \cdots \sum_{A_J \in P_{\mathbb{R}}^J} \min_{z \in \mathcal{Z}_{(k,k')}} \sum_{j=1}^{J} \mathbb{P}\left(Y \in A_j, D = d_j | Z = z\right) \geq 1, \tag{B.9}$$

where the infimum is taken over all partitions $\{P_{\mathbb{R}}^1, \ldots, P_{\mathbb{R}}^J\}$. Next, for all $A_1, \ldots, A_J \in \mathcal{B}_{\mathbb{R}}$,

$$\begin{aligned}
&\mathbb{P}\left(Y_{d_j}(z_k, z_{k'}) \in A_j\right) \\
&= \sum_{A_1 \in P_{\mathbb{R}}^1} \cdots \sum_{A_{j-1} \in P_{\mathbb{R}}^{j-1}} \sum_{A_{j+1} \in P_{\mathbb{R}}^{j+1}} \cdots \sum_{A_J \in P_{\mathbb{R}}^J} \mathbb{P}\left(Y_{d_1}(z_k, z_{k'}) \in A_1, \ldots, Y_{d_J}(z_k, z_{k'}) \in A_J\right) \\
&\leq \sum_{A_1 \in P_{\mathbb{R}}^1} \cdots \sum_{A_{j-1} \in P_{\mathbb{R}}^{j-1}} \sum_{A_{j+1} \in P_{\mathbb{R}}^{j+1}} \cdots \sum_{A_J \in P_{\mathbb{R}}^J} \min_{z \in \mathcal{Z}_{(k,k')}} \sum_{\xi=1}^{J} \mathbb{P}\left(Y \in A_\xi, D = d_\xi | Z = z\right),
\end{aligned}$$

which, together with (B.7), implies the third inequality of Kédagni and Mourifié (2020):

$$\sup_{\{P_{\mathbb{R}}^1, \ldots, P_{\mathbb{R}}^J\}} \max_{j \in \{1, \ldots, J\}} \sup_{A_j \in \mathcal{B}_{\mathbb{R}}} \left\{ \int_{A_j} \max_{z \in \mathcal{Z}_{(k,k')}} f_{Y,D}\left(y, d_j | z\right) \mathrm{d}y - \varphi_j\left(A_j, \mathcal{Z}_{(k,k')}, P_{\mathbb{R}}^1, \ldots, P_{\mathbb{R}}^J\right) \right\} \leq 0, \tag{B.10}$$

where

$$\varphi_j(A_j, \mathcal{W}, P_{\mathbb{R}}^1, \ldots, P_{\mathbb{R}}^J)$$
$$= \sum_{A_1 \in P_{\mathbb{R}}^1} \cdots \sum_{A_{j-1} \in P_{\mathbb{R}}^{j-1}} \sum_{A_{j+1} \in P_{\mathbb{R}}^{j+1}} \cdots \sum_{A_J \in P_{\mathbb{R}}^J} \min_{z \in \mathcal{W}} \sum_{\xi=1}^{J} \int_{A_\xi} f_{Y,D}\,(y, d_\xi | z)\,\mathrm{d}y$$

for all $\mathcal{W} \subset \mathcal{Z}$.

The following lemma shows that when the treatment D is binary, the conditions in (B.8)–(B.10) are weaker than those in (4.1).

Lemma B.1 *If the treatment $D \in \{0,1\}$, then for every $(z_k, z_{k'}) \in \mathscr{Z}$, the restrictions* (B.8)–(B.10) *are implied by those in* (4.1).

Proof of Lemma B.1. Proposition 1.1 of Kitagawa (2015) and Theorem 1 of Mourifié and Wan (2017) show that when both D and Z are binary, the restrictions in (4.1) are sharp for the validity assumption of Z (Assumption 2.1 with $Z \in \{0,1\}$). Suppose that $\mathcal{Z} = \{z_1, \ldots, z_K\}$ and there is some distribution of (Y, D, Z) that satisfies the restrictions in (4.1) for some $(z_k, z_{k'}) \in \mathscr{Z}$, but does not satisfy the restrictions in (B.8)–(B.10) for $(z_k, z_{k'})$. Then we can construct a distribution of (Y', D', Z') with $D' \in \{0,1\}$ and $Z' \in \{0,1\}$ such that for every Borel set A and each $d \in \{0,1\}$,

$$\mathbb{P}(Y' \in A, D' = d, Z' = 0) = \mathbb{P}(Y \in A, D = d | Z = z_k) \cdot \frac{\mathbb{P}(Z = z_k)}{\mathbb{P}(Z = z_k) + \mathbb{P}(Z = z_{k'})} \text{ and}$$
$$\mathbb{P}(Y' \in A, D' = d, Z' = 1) = \mathbb{P}(Y \in A, D = d | Z = z_{k'}) \cdot \frac{\mathbb{P}(Z = z_{k'})}{\mathbb{P}(Z = z_k) + \mathbb{P}(Z = z_{k'})}.$$

Then it can be shown that

$$\mathbb{P}(Z' = 0) = \frac{\mathbb{P}(Z = z_k)}{\mathbb{P}(Z = z_k) + \mathbb{P}(Z = z_{k'})}, \mathbb{P}(Z' = 1) = \frac{\mathbb{P}(Z = z_{k'})}{\mathbb{P}(Z = z_k) + \mathbb{P}(Z = z_{k'})},$$
$$\mathbb{P}(Y' \in A, D' = d | Z' = 0) = \mathbb{P}(Y \in A, D = d | Z = z_k), \text{ and}$$
$$\mathbb{P}(Y' \in A, D' = d | Z' = 1) = \mathbb{P}(Y \in A, D = d | Z = z_{k'}).$$

Since by assumption, the distribution of (Y, D, Z) satisfies the restrictions in (4.1) for $(z_k, z_{k'})$, but does not satisfy the restrictions in (B.8)–(B.10), then $\mathbb{P}(Y' \in A, D' = d | Z' = 0)$ and $\mathbb{P}(Y' \in A, D' = d | Z' = 1)$ satisfy the restrictions in (4.1) with $(z_{k_m}, z_{k'_m})$ replaced by $(0,1)$, but do not satisfy the restrictions in (B.8)–(B.10) with $(z_k, z_{k'})$ replaced by $(0,1)$. This contradicts the sharpness of the restrictions in (4.1). ■

B.4 Definition and Estimation of $\mathscr{Z}_0$

We estimate $\mathscr{Z}_0 = \mathscr{Z}_1 \cap \mathscr{Z}_2$ as $\widehat{\mathscr{Z}_0} = \widehat{\mathscr{Z}_1} \cap \widehat{\mathscr{Z}_2}$, where $\widehat{\mathscr{Z}_1}$ and $\widehat{\mathscr{Z}_2}$ are estimators of $\mathscr{Z}_1$ and $\mathscr{Z}_2$, respectively.

B.4.1 Definition and Estimation of $\mathscr{Z}_1$

The testable implications proposed by Sun (2021) are for full IV validity. Here we extend them to pairwise valid instruments (Definition A.1). We follow the notation of Sun (2021) to introduce the definition of $\mathscr{Z}_1$ and the corresponding estimator. Define conditional probabilities

$$P_z(B, C) = \mathbb{P}(Y \in B, D \in C | Z = z)$$

for all Borel sets $B, C \in \mathcal{B}_{\mathbb{R}}$ and all $z \in \mathcal{Z}$. The testable implications proposed by Sun (2021) for the conditions in Definition A.1 are that for every $m \in \{1, \ldots, \bar{M}\}$,

$$P_{z_{k_m}}(B, \{d_J\}) \leq P_{z_{k'_m}}(B, \{d_J\}) \text{ and } P_{z_{k_m}}(B, \{d_1\}) \geq P_{z_{k'_m}}(B, \{d_1\}) \tag{B.11}$$

for all $B \in \mathcal{B}_{\mathbb{R}}$, and

$$P_{z_{k_m}}(\mathbb{R}, C) \geq P_{z_{k'_m}}(\mathbb{R}, C) \tag{B.12}$$

for all $C = (-\infty, c]$ with $c \in \mathbb{R}$. Without loss of generality, we assume that $d_1 = 0$ and $d_J = 1$. By definition, for all $B, C \in \mathcal{B}_{\mathbb{R}}$,

$$\mathbb{P}(Y \in B, D \in C | Z = z) = \frac{\mathbb{P}(Y \in B, D \in C, Z = z)}{\mathbb{P}(Z = z)}.$$

Next, we reformulate the testable restrictions to define $\mathscr{Z}_1$ and its estimator. Define the following function spaces

$$\begin{aligned}
&\mathcal{G}_P = \left\{ \left(1_{\mathbb{R}\times\mathbb{R}\times\{z_k\}}, 1_{\mathbb{R}\times\mathbb{R}\times\{z_{k'}\}}\right) : k, k' \in \{1, \ldots, K\}, k \neq k' \right\}, \\
&\mathcal{H}_1 = \left\{ (-1)^d \cdot 1_{B\times\{d\}\times\mathbb{R}} : B \text{ is a closed interval in } \mathbb{R}, d \in \{0, 1\} \right\}, \\
&\bar{\mathcal{H}}_1 = \left\{ (-1)^d \cdot 1_{B\times\{d\}\times\mathbb{R}} : B \text{ is a closed, open, or half-closed interval in } \mathbb{R}, d \in \{0, 1\} \right\}, \\
&\mathcal{H}_2 = \left\{ 1_{\mathbb{R}\times C\times\mathbb{R}} : C = (-\infty, c], c \in \mathbb{R} \right\}, \\
&\bar{\mathcal{H}}_2 = \left\{ 1_{\mathbb{R}\times C\times\mathbb{R}} : C = (-\infty, c] \text{ or } C = (-\infty, c), c \in \mathbb{R} \right\}, \\
&\mathcal{H} = \mathcal{H}_1 \cup \mathcal{H}_2, \text{ and } \bar{\mathcal{H}} = \bar{\mathcal{H}}_1 \cup \bar{\mathcal{H}}_2.
\end{aligned} \tag{B.13}$$

Let P and $\widehat{P}$ be defined as in Section 4. Let ϕ, σ^2, $\widehat{\phi}$, and $\widehat{\sigma}^2$ be defined in a way similar to that in Section 4 but for all $(h,g) \in \bar{\mathcal{H}} \times \mathcal{G}_P$ in (B.13). Also, we let $\Lambda(P) = \prod_{k=1}^{K} P\left(1_{\mathbb{R}\times\mathbb{R}\times\{z_k\}}\right)$ and $T_n = n \cdot \prod_{k=1}^{K} \widehat{P}(1_{\mathbb{R}\times\mathbb{R}\times\{z_k\}})$. By similar proof of Lemma 3.1 in Sun (2021), σ^2 and $\widehat{\sigma}^2$ are uniformly bounded in $(h,g) \in \bar{\mathcal{H}} \times \mathcal{G}_P$.

The following lemma reformulates the testable restrictions in terms of ϕ.

Lemma B.2 *Suppose that the instrument Z is pairwise valid for the treatment D with the largest validity pair set $\mathscr{Z}_{\bar{M}} = \{(z_{k_1}, z_{k'_1}), \ldots, (z_{k_{\bar{M}}}, z_{k'_{\bar{M}}})\}$. For every $m \in \{1, \ldots, \bar{M}\}$, we have that $\sup_{h\in\mathcal{H}} \phi(h,g) = 0$ with $g = (1_{\mathbb{R}\times\mathbb{R}\times\{z_{k_m}\}}, 1_{\mathbb{R}\times\mathbb{R}\times\{z_{k'_m}\}})$.*

Proof of Lemma B.2. Note that for every $g \in \mathcal{G}_P$, we can always find some $a \in \mathbb{R}$ such that $\phi(h,g) = 0$ with $h = 1_{\{a\}\times\{0\}\times\mathbb{R}}$. So $\sup_{h\in\mathcal{H}} \phi(h,g) \geq 0$ for every $g \in \mathcal{G}_P$. Under assumption, for every $g = (1_{\mathbb{R}\times\mathbb{R}\times\{z_{k_m}\}}, 1_{\mathbb{R}\times\mathbb{R}\times\{z_{k'_m}\}})$, by Lemma 2.1 of Sun (2021), $\phi(h,g) \leq 0$ for all $h \in \mathcal{H}$. Thus, $\sup_{h\in\mathcal{H}} \phi(h,g) = 0$. ■

Lemma B.2 provides a necessary condition for $\mathscr{Z}_{\bar{M}}$. By Lemma B.2, we define

$$\mathcal{G}_1 = \left\{ g \in \mathcal{G}_P : \sup_{h\in\mathcal{H}} \phi(h,g) = 0 \right\} \text{ and } \widehat{\mathcal{G}}_1 = \left\{ g \in \mathcal{G}_P : \sqrt{T_n} \left| \sup_{h\in\mathcal{H}} \frac{\widehat{\phi}(h,g)}{\xi_0 \vee \widehat{\sigma}(h,g)} \right| \leq \tau_n \right\} \tag{B.14}$$

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 define $\mathscr{Z}_1$ as the collection of all (z,z') that are associated with some $g \in \mathcal{G}_1$:

$$\mathscr{Z}_1 = \left\{ (z_k, z_{k'}) \in \mathscr{Z} : g = (1_{\mathbb{R}\times\mathbb{R}\times\{z_k\}}, 1_{\mathbb{R}\times\mathbb{R}\times\{z_{k'}\}}) \in \mathcal{G}_1 \right\}. \tag{B.15}$$

We use $\widehat{\mathcal{G}}_1$ to construct the estimator of $\mathscr{Z}_1$, denoted by $\widehat{\mathscr{Z}_1}$, which is defined as the set of all (z,z') that are associated with some $g \in \widehat{\mathcal{G}}_1$ in the same way $\mathscr{Z}_1$ is defined based on $\mathcal{G}_1$:

$$\widehat{\mathscr{Z}_1} = \left\{ (z_k, z_{k'}) \in \mathscr{Z} : g = (1_{\mathbb{R}\times\mathbb{R}\times\{z_k\}}, 1_{\mathbb{R}\times\mathbb{R}\times\{z_{k'}\}}) \in \widehat{\mathcal{G}}_1 \right\}. \tag{B.16}$$

To establish consistency of $\widehat{\mathscr{Z}_1}$, we state and prove an auxiliary lemma.

Lemma B.3 *Under Assumption A.2, $\widehat{\phi} \to \phi$, $T_n/n \to \Lambda(P)$, and $\widehat{\sigma} \to \sigma$ almost uniformly.*[18] *In addition, $\sqrt{T_n}(\widehat{\phi} - \phi) \rightsquigarrow \mathbb{G}$ for some random element $\mathbb{G}$, and for all $(h,g) \in \bar{\mathcal{H}} \times \mathcal{G}_P$ with $g = (g_1, g_2)$, the variance $Var(\mathbb{G}(h,g)) = \sigma^2(h,g)$.*

Proof of Lemma B.3. Note that the $\mathcal{G}_P$ defined in (B.13) is only slightly different from

[18] See the definition of almost uniform convergence in van der Vaart and Wellner (1996, p. 52).

the $\mathcal{G}$ defined in (7) of Sun (2021). The lemma can be proved following a strategy similar to that of the proofs of Lemmas C.11 and 3.1 of Sun (2021). ■

The following proposition establishes consistency of $\widehat{\mathscr{X}_1}$.

Proposition B.1 *Under Assumption A.2,* $\mathbb{P}(\widehat{\mathcal{G}}_1 = \mathcal{G}_1) \to 1$*, and thus* $\mathbb{P}(\widehat{\mathscr{X}_1} = \mathscr{X}_1) \to 1$.

Proof of Proposition B.1. First, suppose $\mathcal{G}_1 \neq \varnothing$. Under the constructions, we have that for all $\varepsilon > 0$,

$$\begin{aligned}
&\lim_{n\to\infty} \mathbb{P}\left(\mathcal{G}_1 \setminus \widehat{\mathcal{G}}_1 \neq \varnothing\right) \\
&\leq \lim_{n\to\infty} \mathbb{P}\left(\max_{g\in\mathcal{G}_1} \sqrt{T_n} \left| \sup_{h\in\mathcal{H}} \left(\frac{\widehat{\phi}(h,g)}{\xi_0 \vee \widehat{\sigma}(h,g)} \right) - \sup_{h\in\mathcal{H}} \left(\frac{\phi(h,g)}{\xi_0 \vee \widehat{\sigma}(h,g)} \right) \right| > \tau_n \right) \\
&\leq \lim_{n\to\infty} \mathbb{P}\left(\max_{g\in\mathcal{G}_1} \sup_{h\in\mathcal{H}} \sqrt{T_n} \left| \frac{\widehat{\phi}(h,g) - \phi(h,g)}{\xi_0 \vee \widehat{\sigma}(h,g)} \right| > \tau_n \right).
\end{aligned}$$

By Lemma B.3, $\sqrt{T_n}(\widehat{\phi} - \phi) \rightsquigarrow \mathbb{G}$ and $\widehat{\sigma} \to \sigma$ almost uniformly, which implies that $\widehat{\sigma} \rightsquigarrow \sigma$ 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),

$$\max_{g\in\mathcal{G}_1} \sup_{h\in\mathcal{H}} \sqrt{T_n} \left| \frac{\widehat{\phi}(h,g) - \phi(h,g)}{\xi_0 \vee \widehat{\sigma}(h,g)} \right| \rightsquigarrow \max_{g\in\mathcal{G}_1} \sup_{h\in\mathcal{H}} \left| \frac{\mathbb{G}(h,g)}{\xi_0 \vee \sigma(h,g)} \right|.$$

Since $\tau_n \to \infty$, we have that $\lim_{n\to\infty} \mathbb{P}(\mathcal{G}_1 \setminus \widehat{\mathcal{G}}_1 \neq \varnothing) = 0$.

If $\mathcal{G}_1 = \mathcal{G}_P$, then clearly $\lim_{n\to\infty} \mathbb{P}(\widehat{\mathcal{G}}_1 \setminus \mathcal{G}_1 \neq \varnothing) = 0$. Suppose $\mathcal{G}_1 \neq \mathcal{G}_P$. Since $\mathcal{G}_P$ is a finite set and $\widehat{\sigma}$ is uniformly bounded in (h,g) by construction, then there is a $\delta > 0$ such that $\min_{g\in\mathcal{G}_P\setminus\mathcal{G}_1} |\sup_{h\in\mathcal{H}} \phi(h,g) / (\xi_0 \vee \widehat{\sigma}(h,g))| > \delta$. Thus, we have that

$$\begin{aligned}
&\lim_{n\to\infty} \mathbb{P}\left(\widehat{\mathcal{G}}_1 \setminus \mathcal{G}_1 \neq \varnothing\right) \\
&\leq \lim_{n\to\infty} \mathbb{P}\left(\max_{g\in\widehat{\mathcal{G}}_1\setminus\mathcal{G}_1} \left| \sup_{h\in\mathcal{H}} \frac{\phi(h,g)}{\xi_0 \vee \widehat{\sigma}(h,g)} \right| > \delta, \max_{g\in\widehat{\mathcal{G}}_1\setminus\mathcal{G}_1} \sqrt{T_n} \left| \sup_{h\in\mathcal{H}} \frac{\widehat{\phi}(h,g)}{\xi_0 \vee \widehat{\sigma}(h,g)} \right| \leq \tau_n \right).
\end{aligned}$$

By Lemma B.3, $\widehat{\phi} \to \phi$ almost uniformly. Thus, for every $\varepsilon > 0$, there is a measurable set A with $\mathbb{P}(A) \geq 1 - \varepsilon$ such that for sufficiently large n,

$$\max_{g\in\mathcal{G}_P} \left| \left| \sup_{h\in\mathcal{H}} \frac{\widehat{\phi}(h,g)}{\xi_0 \vee \widehat{\sigma}(h,g)} \right| - \left| \sup_{h\in\mathcal{H}} \frac{\phi(h,g)}{\xi_0 \vee \widehat{\sigma}(h,g)} \right| \right| \leq \frac{\delta}{2}$$

uniformly on A. We now have that

$$\begin{aligned}
&\lim_{n\to\infty}\mathbb{P}\left(\widehat{\mathcal{G}}_1 \setminus \mathcal{G}_1 \neq \varnothing\right) \\
&\leq \lim_{n\to\infty}\mathbb{P}\left(\begin{array}{c}\left\{\max_{g\in\widehat{\mathcal{G}}_1\setminus\mathcal{G}_1}\left|\sup_{h\in\mathcal{H}}\frac{\phi(h,g)}{\xi_0\vee\widehat{\sigma}(h,g)}\right| > \delta\right\} \\ \cap\left\{\max_{g\in\widehat{\mathcal{G}}_1\setminus\mathcal{G}_1}\sqrt{T_n}\left|\sup_{h\in\mathcal{H}}\frac{\widehat{\phi}(h,g)}{\xi_0\vee\widehat{\sigma}(h,g)}\right| \leq \tau_n\right\}\cap A\end{array}\right) + \mathbb{P}(A^c) \\
&\leq \lim_{n\to\infty}\mathbb{P}\left(\sqrt{\frac{T_n}{n}}\frac{\delta}{2} < \max_{g\in\widehat{\mathcal{G}}_1\setminus\mathcal{G}_1}\sqrt{\frac{T_n}{n}}\left|\sup_{h\in\mathcal{H}}\frac{\widehat{\phi}(h,g)}{\xi_0\vee\widehat{\sigma}(h,g)}\right| \leq \frac{\tau_n}{\sqrt{n}}\right) + \varepsilon = \varepsilon,
\end{aligned}$$

because $\tau_n/\sqrt{n} \to 0$ as $n \to \infty$. Here, ε can be arbitrarily small. Thus we have that $\mathbb{P}(\widehat{\mathcal{G}}_1 = \mathcal{G}_1) \to 1$, because $\mathbb{P}(\mathcal{G}_1\setminus\widehat{\mathcal{G}}_1 \neq \varnothing) \to 0$ and $\mathbb{P}(\widehat{\mathcal{G}}_1 \setminus \mathcal{G}_1 \neq \varnothing) \to 0$.

Second, suppose $\mathcal{G}_1 = \varnothing$. This implies that $\min_{g\in\mathcal{G}_P}\left|\sup_{h\in\mathcal{H}}\phi(h,g)/(\xi_0\vee\widehat{\sigma}(h,g))\right| > \delta$ for some $\delta > 0$. Since by Lemma B.3, $\widehat{\phi} \to \phi$ almost uniformly, then there is a measurable set A with $\mathbb{P}(A) \geq 1-\varepsilon$ such that for sufficiently large n,

$$\max_{g\in\mathcal{G}_P}\left|\left|\sup_{h\in\mathcal{H}}\frac{\widehat{\phi}(h,g)}{\xi_0\vee\widehat{\sigma}(h,g)}\right| - \left|\sup_{h\in\mathcal{H}}\frac{\phi(h,g)}{\xi_0\vee\widehat{\sigma}(h,g)}\right|\right| \leq \frac{\delta}{2}$$

uniformly on A. Thus we now have that

$$\begin{aligned}
\lim_{n\to\infty}\mathbb{P}\left(\widehat{\mathcal{G}}_1 \neq \varnothing\right) &\leq \lim_{n\to\infty}\mathbb{P}\left(\begin{array}{c}\left\{\max_{g\in\widehat{\mathcal{G}}_1}\left|\sup_{h\in\mathcal{H}}\frac{\phi(h,g)}{\xi_0\vee\widehat{\sigma}(h,g)}\right| > \delta\right\} \\ \cap\left\{\max_{g\in\widehat{\mathcal{G}}_1}\sqrt{T_n}\left|\sup_{h\in\mathcal{H}}\frac{\widehat{\phi}(h,g)}{\xi_0\vee\widehat{\sigma}(h,g)}\right| \leq \tau_n\right\}\cap A\end{array}\right) + \mathbb{P}(A^c) \\
&\leq \lim_{n\to\infty}\mathbb{P}\left(\sqrt{\frac{T_n}{n}}\frac{\delta}{2} < \max_{g\in\widehat{\mathcal{G}}_1}\sqrt{\frac{T_n}{n}}\left|\sup_{h\in\mathcal{H}}\frac{\widehat{\phi}(h,g)}{\xi_0\vee\widehat{\sigma}(h,g)}\right| \leq \frac{\tau_n}{\sqrt{n}}\right) + \varepsilon = \varepsilon,
\end{aligned}$$

because $\tau_n/\sqrt{n} \to 0$ as $n \to \infty$. Here, ε can be arbitrarily small. Thus, $\mathbb{P}(\widehat{\mathcal{G}}_1 = \mathcal{G}_1) = 1 - \mathbb{P}(\widehat{\mathcal{G}}_1 \neq \varnothing) \to 1$. ■

As mentioned after Proposition 4.1, Proposition B.1 is related to the contact set estimation in Sun (2021). Since $\mathcal{G}_1 \subset \mathcal{G}_P$ and $\mathcal{G}_P$ is a finite set, we can use techniques similar to those in Sun (2021) to obtain the stronger result in Proposition B.1, that is, $\mathbb{P}(\widehat{\mathcal{G}}_1 = \mathcal{G}_1) \to 1$.

B.4.2 Definition and Estimation of $\mathscr{Z}_2$

The definition of $\mathscr{Z}_2$ relies on the testable implications in Kédagni and Mourifié (2020). Under Conditions (i) and (ii) in Definition A.1, we can define $Y_d(z,z')$ for every $d \in \mathcal{D}$ and

every $(z, z') \in \mathscr{Z}_{\bar{M}}$ such that $Y_d(z, z') = Y_{dz} = Y_{dz'}$ a.s. We consider the case where Y is continuous. Similar results can be obtained easily when Y is discrete. To avoid theoretical and computational complications, we introduce the following testable implications that are slightly weaker than (and implied by) the original testable restrictions in Kédagni and Mourifié (2020) (see Appendix B.3).

Let $\mathcal{R}$ denote the collection of all subsets $C \subset \mathbb{R}$ such that $C = (a, b]$ with $a, b \in \mathbb{R}$ and $a < b$. For every $\mathcal{Z}_{(k,k')} = (z_k, z_{k'}) \in \mathscr{Z}_{\bar{M}}$, every $A \in \mathcal{B}_{\mathbb{R}}$, every $d \in \mathcal{D}$, and each $z \in \mathcal{Z}_{(k,k')}$,

$$\mathbb{P}\left(Y \in A, D = d | Z = z\right) \leq \mathbb{P}\left(Y_{dz} \in A | Z = z\right) = \mathbb{P}\left(Y_d(z_k, z_{k'}) \in A\right),$$

which implies that

$$\max_{z \in \mathcal{Z}_{(k,k')}} \mathbb{P}\left(Y \in A, D = d | Z = z\right) \leq \mathbb{P}\left(Y_d(z_k, z_{k'}) \in A\right). \tag{B.17}$$

Let $\mathscr{P}$ be a prespecified finite collection of partitions $P_{\mathbb{R}}$ of $\mathbb{R}$ such that $P_{\mathbb{R}} = \{C_1, \ldots, C_N\}$ for some N with $C_k \in \mathcal{R}$ for all k, $\cup_{k=1}^{N} C_k = \mathbb{R}$, and $C_k \cap C_l = \varnothing$ for all $k \neq l$. Then we obtain the first condition:

$$\max_{P_{\mathbb{R}} \in \mathscr{P}} \max_{d \in \mathcal{D}} \sum_{A \in P_{\mathbb{R}}} \max_{z \in \mathcal{Z}_{(k,k')}} \mathbb{P}\left(Y \in A, D = d | Z = z\right) \leq \max_{P_{\mathbb{R}} \in \mathscr{P}} \max_{d \in \mathcal{D}} \sum_{A \in P_{\mathbb{R}}} \mathbb{P}\left(Y_d\left(z_k, z_{k'}\right) \in A\right) = 1. \tag{B.18}$$

Also, for all $A_1, \ldots, A_J \in \mathcal{B}_{\mathbb{R}}$,

$$\begin{aligned}
&\mathbb{P}\left(Y_{d_1}(z_k, z_{k'}) \in A_1, \ldots, Y_{d_J}(z_k, z_{k'}) \in A_J\right) \\
&= \min_{z \in \mathcal{Z}_{(k,k')}} \mathbb{P}\left(Y_{d_1}(z_k, z_{k'}) \in A_1, \ldots, Y_{d_J}(z_k, z_{k'}) \in A_J | Z = z\right) \\
&= \min_{z \in \mathcal{Z}_{(k,k')}} \sum_{j=1}^{J} \mathbb{P}\left(Y_{d_1}(z_k, z_{k'}) \in A_1, \ldots, Y_{d_J}(z_k, z_{k'}) \in A_J, D = d_j | Z = z\right) \\
&\leq \min_{z \in \mathcal{Z}_{(k,k')}} \sum_{j=1}^{J} \mathbb{P}\left(Y \in A_j, D = d_j | Z = z\right).
\end{aligned}$$

Let $P^1_{\mathbb{R}}, \ldots, P^J_{\mathbb{R}} \in \mathscr{P}$. It follows that

$$\begin{aligned} 1 &= \sum_{A_1 \in P^1_{\mathbb{R}}} \cdots \sum_{A_J \in P^J_{\mathbb{R}}} \mathbb{P}\left(Y_{d_1}(z_k, z_{k'}) \in A_1, \ldots, Y_{d_J}(z_k, z_{k'}) \in A_J\right) \\ &\leq \sum_{A_1 \in P^1_{\mathbb{R}}} \cdots \sum_{A_J \in P^J_{\mathbb{R}}} \min_{z \in \mathcal{Z}_{(k,k')}} \sum_{j=1}^{J} \mathbb{P}\left(Y \in A_j, D = d_j | Z = z\right). \end{aligned}$$

Then we obtain the second condition:

$$\min_{P^1_{\mathbb{R}}, \ldots, P^J_{\mathbb{R}} \in \mathscr{P}} \sum_{A_1 \in P^1_{\mathbb{R}}} \cdots \sum_{A_J \in P^J_{\mathbb{R}}} \min_{z \in \mathcal{Z}_{(k,k')}} \sum_{j=1}^{J} \mathbb{P}\left(Y \in A_j, D = d_j | Z = z\right) \geq 1. \tag{B.19}$$

Next, for every j and every $A_j \in \mathcal{B}_{\mathbb{R}}$,

$$\begin{aligned} &\mathbb{P}\left(Y_{d_j}(z_k, z_{k'}) \in A_j\right) \\ &= \sum_{A_1 \in P^1_{\mathbb{R}}} \cdots \sum_{A_{j-1} \in P^{j-1}_{\mathbb{R}}} \sum_{A_{j+1} \in P^{j+1}_{\mathbb{R}}} \cdots \sum_{A_J \in P^J_{\mathbb{R}}} \mathbb{P}\left(Y_{d_1}(z_k, z_{k'}) \in A_1, \ldots, Y_{d_J}(z_k, z_{k'}) \in A_J\right) \\ &\leq \sum_{A_1 \in P^1_{\mathbb{R}}} \cdots \sum_{A_{j-1} \in P^{j-1}_{\mathbb{R}}} \sum_{A_{j+1} \in P^{j+1}_{\mathbb{R}}} \cdots \sum_{A_J \in P^J_{\mathbb{R}}} \min_{z \in \mathcal{Z}_{(k,k')}} \sum_{\xi=1}^{J} \mathbb{P}\left(Y \in A_\xi, D = d_\xi | Z = z\right), \end{aligned}$$

which, together with (B.17), implies the third condition:

$$\begin{aligned} \max_{P^1_{\mathbb{R}}, \ldots, P^J_{\mathbb{R}} \in \mathscr{P}} \max_{j \in \{1, \ldots, J\}} \sup_{A_j \in \mathcal{R}} \Bigg\{ &\max_{z \in \mathcal{Z}_{(k,k')}} \mathbb{P}(Y \in A_j, D = d_j | Z = z) \\ &- \varphi_j\left(A_j, \mathcal{Z}_{(k,k')}, P^1_{\mathbb{R}}, \ldots, P^J_{\mathbb{R}}\right) \Bigg\} \leq 0, \end{aligned} \tag{B.20}$$

where

$$\begin{aligned} &\varphi_j\left(A_j, \mathcal{W}, P^1_{\mathbb{R}}, \ldots, P^J_{\mathbb{R}}\right) \\ &= \sum_{A_1 \in P^1_{\mathbb{R}}} \cdots \sum_{A_{j-1} \in P^{j-1}_{\mathbb{R}}} \sum_{A_{j+1} \in P^{j+1}_{\mathbb{R}}} \cdots \sum_{A_J \in P^J_{\mathbb{R}}} \min_{z \in \mathcal{W}} \sum_{\xi=1}^{J} \mathbb{P}\left(Y \in A_\xi, D = d_\xi | Z = z\right) \end{aligned}$$

for all $\mathcal{W} \subset \mathcal{Z}$.

Next, we reformulate the testable implications in (B.18)–(B.20) to define $\mathscr{Z}_2$ and $\widehat{\mathscr{Z}_2}$.

Define the function spaces

$$\mathcal{G}_Z = \left\{1_{\mathbb{R}\times\mathbb{R}\times\{z_k\}} : 1 \le k \le K\right\}, \mathcal{H}_D = \{1_{\mathbb{R}\times\{d\}\times\mathbb{R}}, d \in \mathcal{D}\}, \mathcal{H}_B = \{1_{B\times\mathbb{R}\times\mathbb{R}} : B \in \mathcal{R}\},$$
$$\text{and } \bar{\mathcal{H}}_B = \{1_{B\times\mathbb{R}\times\mathbb{R}} : B \text{ is a closed, open, or half-closed interval in } \mathbb{R}\}. \tag{B.21}$$

Let P and $\widehat{P}$ be defined as in Section 4. Define a map $\psi : \bar{\mathcal{H}}_B \times \mathcal{H}_D \times \mathcal{G}_Z \to \mathbb{R}$ such that

$$\psi(h, f, g) = \frac{P(h \cdot f \cdot g)}{P(g)}$$

for every $(h, f, g) \in \bar{\mathcal{H}}_B \times \mathcal{H}_D \times \mathcal{G}_Z$. Moreover, define a map $\mathbb{H}$ such that if $P_\mathbb{R} \in \mathscr{P}$ with $P_\mathbb{R} = \{C_1, \ldots, C_N\}$ and $C_k \in \mathcal{R}$ for all $k \in \{1, \ldots, N\}$, then

$$\mathbb{H}(P_\mathbb{R}) = \{1_{C\times\mathbb{R}\times\mathbb{R}} : C \in P_\mathbb{R}\}. \tag{B.22}$$

Let $\mathrm{P}(\mathcal{G}_Z)$ be the collection of all nonempty subsets of $\mathcal{G}_Z$. Then for every $\mathcal{G}_S \in \mathrm{P}(\mathcal{G}_Z)$, define

$$\psi_1(\mathcal{G}_S) = \max_{P_\mathbb{R}\in\mathscr{P}} \max_{f\in\mathcal{H}_D} \sum_{h\in\mathbb{H}(P_\mathbb{R})} \max_{g\in\mathcal{G}_S} \psi(h, f, g) - 1,$$

$$\psi_2(\mathcal{G}_S) = 1 - \min_{P^1_\mathbb{R},\ldots,P^J_\mathbb{R}\in\mathscr{P}} \sum_{h_1\in\mathbb{H}\left(P^1_\mathbb{R}\right)} \cdots \sum_{h_J\in\mathbb{H}\left(P^J_\mathbb{R}\right)} \min_{g\in\mathcal{G}_S} \sum_{j=1}^{J} \psi(h_j, f_j, g),$$

and

$$\psi_3(\mathcal{G}_S) = \max_{P^1_\mathbb{R},\ldots,P^J_\mathbb{R}\in\mathscr{P}} \max_{j\in\{1,\ldots,J\}} \sup_{h_j\in\mathcal{H}_B} \left\{\max_{g\in\mathcal{G}_S} \psi(h_j, f_j, g) - \tilde{\varphi}_j\left(h_j, \mathcal{G}_S, P^1_\mathbb{R}, \ldots, P^J_\mathbb{R}\right)\right\},$$

where $f_j = 1_{\mathbb{R}\times\{d_j\}\times\mathbb{R}}$ and

$$\begin{aligned}&\tilde{\varphi}_j\left(h_j, \mathcal{G}_S, P^1_\mathbb{R}, \ldots, P^J_\mathbb{R}\right)\\ &= \sum_{h_1\in\mathbb{H}\left(P^1_\mathbb{R}\right)} \cdots \sum_{h_{j-1}\in\mathbb{H}\left(P^{j-1}_\mathbb{R}\right)} \sum_{h_{j+1}\in\mathbb{H}\left(P^{j+1}_\mathbb{R}\right)} \cdots \sum_{h_J\in\mathbb{H}\left(P^J_\mathbb{R}\right)} \min_{g\in\mathcal{G}_S} \sum_{\xi=1}^{J} \psi(h_\xi, f_\xi, g).\end{aligned}$$

For every $\mathcal{Z}_{(k,k')} \in \mathscr{Z}_{\bar{M}}$, let $\mathcal{G}(\mathcal{Z}_{(k,k')}) = \{1_{\mathbb{R}\times\mathbb{R}\times\{z_k\}}, 1_{\mathbb{R}\times\mathbb{R}\times\{z_{k'}\}}\}$. The conditions in (B.18)–(B.20) imply that $\psi_l(\mathcal{G}(\mathcal{Z}_{(k,k')})) \le 0$ for all $l \in \{1, 2, 3\}$. Thus, we define $\mathscr{Z}_2$ by

$$\mathscr{Z}_2 = \left\{\mathcal{Z}_{(k,k')} \in \mathscr{Z} : \psi_l(\mathcal{G}(\mathcal{Z}_{(k,k')})) \le 0, l \in \{1, 2, 3\}\right\}.$$

Let $\widehat{\psi} : \bar{\mathcal{H}}_B \times \mathcal{H}_D \times \mathcal{G}_Z \to \mathbb{R}$ be the sample analog of ψ such that

$$\widehat{\psi}(h, f, g) = \frac{\widehat{P}(h \cdot f \cdot g)}{\widehat{P}(g)}$$

for every $(h, f, g) \in \bar{\mathcal{H}}_B \times \mathcal{H}_D \times \mathcal{G}_Z$. Let $\widehat{\psi}_l$ be the sample analog of ψ_l for $l \in \{1, 2, 3\}$, which replaces ψ in ψ_l by $\widehat{\psi}$. We define the estimator $\widehat{\mathscr{Z}_2}$ for $\mathscr{Z}_2$ by

$$\widehat{\mathscr{Z}_2} = \left\{ \mathcal{Z}_{(k,k')} \in \mathscr{Z} : \sqrt{T_n}\widehat{\psi}_l(\mathcal{G}(\mathcal{Z}_{(k,k')})) \leq t_n, l \in \{1, 2, 3\} \right\},$$

where $T_n = n \cdot \prod_{k=1}^{K} \widehat{P}(1_{\mathbb{R}\times\mathbb{R}\times\{z_k\}})$, $t_n \to \infty$, and $t_n/\sqrt{n} \to 0$ as $n \to \infty$.

To establish consistency of $\widehat{\mathscr{Z}_2}$, we state and prove some auxiliary lemmas.

Lemma B.4 *The function space $\mathcal{H}_B$ is a VC class with VC index $V(\mathcal{H}_B) = 3$.*

Proof of Lemma B.4. The proof closely follows the strategy of the proof of Lemma C.2 of Sun (2021). ■

We define

$$\mathcal{V} = \{h \cdot f \cdot g : h \in \bar{\mathcal{H}}_B, f \in \mathcal{H}_D, g \in \mathcal{G}_Z\} \text{ and } \tilde{\mathcal{V}} = \mathcal{V} \cup \mathcal{G}_Z. \tag{B.23}$$

Lemma B.5 *The function space $\tilde{\mathcal{V}}$ defined in (B.23) is Donsker and pre-Gaussian uniformly in $Q \in \mathcal{P}$, and $\tilde{\mathcal{V}}$ is Glivenko–Cantelli uniformly in $Q \in \mathcal{P}$.*

Proof of Lemma B.5. The proof closely follows the strategies of the proofs of Lemmas C.5 and C.6 of Sun (2021). ■

The following proposition establishes consistency of $\widehat{\mathscr{Z}_2}$.

Proposition B.2 *Under Assumption A.2, $\mathbb{P}(\widehat{\mathscr{Z}_2} = \mathscr{Z}_2) \to 1$.*

Proof of Proposition B.2. Let $\mathcal{C}_2$ be the set of all $\mathcal{G}(\mathcal{Z}_{(k,k')})$ with $\mathcal{Z}_{(k,k')} \in \mathscr{Z}_2$ and $\widehat{\mathcal{C}_2}$ be the

set of all $\mathcal{G}(\mathcal{Z}_{(k,k')})$ with $\mathcal{Z}_{(k,k')} \in \widehat{\mathscr{Z}_2}$. First, we have that

$$\begin{aligned}\mathbb{P}\left(\mathcal{C}_2 \setminus \widehat{\mathcal{C}}_2 \neq \varnothing\right) \leq & \mathbb{P}\left(\max_{\mathcal{G}_S \in \mathcal{C}_2 \setminus \widehat{\mathcal{C}_2}} \sqrt{T_n}\left\{\widehat{\psi}_1\left(\mathcal{G}_S\right) - \psi_1\left(\mathcal{G}_S\right)\right\} > t_n\right) \\ & + \mathbb{P}\left(\max_{\mathcal{G}_S \in \mathcal{C}_2 \setminus \widehat{\mathcal{C}_2}} \sqrt{T_n}\left\{\widehat{\psi}_2\left(\mathcal{G}_S\right) - \psi_2\left(\mathcal{G}_S\right)\right\} > t_n\right) \\ & + \mathbb{P}\left(\max_{\mathcal{G}_S \in \mathcal{C}_2 \setminus \widehat{\mathcal{C}_2}} \sqrt{T_n}\left\{\widehat{\psi}_3\left(\mathcal{G}_S\right) - \psi_3\left(\mathcal{G}_S\right)\right\} > t_n\right).\end{aligned}$$

By Theorem 1.3.6 (continuous mapping) of van der Vaart and Wellner (1996),

$$\begin{aligned}&\max_{\mathcal{G}_S \in \mathcal{C}_2} \sqrt{T_n}\left|\max_{P_{\mathbb{R}} \in \mathscr{P}} \max_{f \in \mathcal{H}_D} \sum_{h \in \mathbb{H}(P_{\mathbb{R}})} \max_{g \in \mathcal{G}_S} \widehat{\psi}\left(h, f, g\right) - \max_{P_{\mathbb{R}} \in \mathscr{P}} \max_{f \in \mathcal{H}_D} \sum_{h \in \mathbb{H}(P_{\mathbb{R}})} \max_{g \in \mathcal{G}_S} \psi\left(h, f, g\right)\right| \\ &\leq \max_{\mathcal{G}_S \in \mathcal{C}_2} \max_{P_{\mathbb{R}} \in \mathscr{P}} \max_{f \in \mathcal{H}_D} \sum_{h \in \mathbb{H}(P_{\mathbb{R}})} \max_{g \in \mathcal{G}_S} \sqrt{T_n}\left|\widehat{\psi}\left(h, f, g\right) - \psi\left(h, f, g\right)\right| \rightsquigarrow \mathbb{G}_1\end{aligned}$$

for some random element $\mathbb{G}_1$. Then it follows that

$$\begin{aligned}\mathbb{P}\left(\max_{\mathcal{G}_S \in \mathcal{C}_2 \setminus \widehat{\mathcal{C}_2}} \sqrt{T_n}\left\{\widehat{\psi}_1\left(\mathcal{G}_S\right) - \psi_1\left(\mathcal{G}_S\right)\right\} > t_n\right) &\leq \mathbb{P}\left(\max_{\mathcal{G}_S \in \mathcal{C}_2} \sqrt{T_n}\left|\widehat{\psi}_1\left(\mathcal{G}_S\right) - \psi_1\left(\mathcal{G}_S\right)\right| > t_n\right) \\ &\to 0.\end{aligned}$$

Similarly, we have that

$$\mathbb{P}\left(\max_{\mathcal{G}_S \in \mathcal{C}_2 \setminus \widehat{\mathcal{C}_2}} \sqrt{T_n}\left\{\widehat{\psi}_2\left(\mathcal{G}_S\right) - \psi_2\left(\mathcal{G}_S\right)\right\} > t_n\right) \to 0$$

and

$$\mathbb{P}\left(\max_{\mathcal{G}_S \in \mathcal{C}_2 \setminus \widehat{\mathcal{C}_2}} \sqrt{T_n}\left\{\widehat{\psi}_3\left(\mathcal{G}_S\right) - \psi_3\left(\mathcal{G}_S\right)\right\} > t_n\right) \to 0.$$

Thus, $\mathbb{P}(\mathcal{C}_2 \setminus \widehat{\mathcal{C}}_2 \neq \varnothing) \to 0$.

Next, let $\mathcal{C}$ be the set of all $\mathcal{G}(\mathcal{Z}_{(k,k')})$ with $\mathcal{Z}_{(k,k')} \in \mathscr{Z}$. Clearly, $\mathcal{C}$ is a finite set. If $\mathcal{C} \setminus \mathcal{C}_2 \neq \varnothing$, there is some $\delta > 0$ such that $\min_{\mathcal{G}_S \in \mathcal{C} \setminus \mathcal{C}_2} \max_{l \in \{1,2,3\}} \psi_l\left(\mathcal{G}_S\right) > \delta$. Then we have

that

$$\begin{aligned}\mathbb{P}\left(\widehat{\mathcal{C}}_2 \setminus \mathcal{C}_2 \neq \varnothing\right) \leq & \mathbb{P}\left(\max_{\mathcal{G}_S \in \widehat{\mathcal{C}_2}\setminus\mathcal{C}_2} \psi_1\left(\mathcal{G}_S\right) > \delta, \max_{\mathcal{G}_S \in \widehat{\mathcal{C}_2}\setminus\mathcal{C}_2} \sqrt{T_n}\widehat{\psi}_1\left(\mathcal{G}_S\right) \leq t_n\right) \\ & + \mathbb{P}\left(\max_{\mathcal{G}_S \in \widehat{\mathcal{C}_2}\setminus\mathcal{C}_2} \psi_2\left(\mathcal{G}_S\right) > \delta, \max_{\mathcal{G}_S \in \widehat{\mathcal{C}_2}\setminus\mathcal{C}_2} \sqrt{T_n}\widehat{\psi}_2\left(\mathcal{G}_S\right) \leq t_n\right) \\ & + \mathbb{P}\left(\max_{\mathcal{G}_S \in \widehat{\mathcal{C}_2}\setminus\mathcal{C}_2} \psi_3\left(\mathcal{G}_S\right) > \delta, \max_{\mathcal{G}_S \in \widehat{\mathcal{C}_2}\setminus\mathcal{C}_2} \sqrt{T_n}\widehat{\psi}_3\left(\mathcal{G}_S\right) \leq t_n\right).\end{aligned}$$

By Lemma B.5 and Lemma 1.9.3 of van der Vaart and Wellner (1996), $\|\widehat{\psi} - \psi\|_\infty \to 0$ almost uniformly. Then we have that

$$\begin{aligned}&\max_{\mathcal{G}_S \in \mathcal{C}} \left|\widehat{\psi}_1\left(\mathcal{G}_S\right) - \psi_1\left(\mathcal{G}_S\right)\right| \\ &= \max_{\mathcal{G}_S \in \mathcal{C}} \left|\max_{P_{\mathbb{R}} \in \mathscr{P}} \max_{f \in \mathcal{H}_D} \sum_{h \in \mathbb{H}(P_{\mathbb{R}})} \max_{g \in \mathcal{G}_S} \widehat{\psi}\left(h, f, g\right) - \max_{P_{\mathbb{R}} \in \mathscr{P}} \max_{f \in \mathcal{H}_D} \sum_{h \in \mathbb{H}(P_{\mathbb{R}})} \max_{g \in \mathcal{G}_S} \psi\left(h, f, g\right)\right| \\ &\leq \max_{\mathcal{G}_S \in \mathcal{C}} \max_{P_{\mathbb{R}} \in \mathscr{P}} \max_{f \in \mathcal{H}_D} \sum_{h \in \mathbb{H}(P_{\mathbb{R}})} \max_{g \in \mathcal{G}_S} \left|\widehat{\psi}\left(h, f, g\right) - \psi\left(h, f, g\right)\right| \to 0\end{aligned}$$

almost uniformly. Similarly, it follows that

$$\max_{\mathcal{G}_S \in \mathcal{C}} \left|\widehat{\psi}_2\left(\mathcal{G}_S\right) - \psi_2\left(\mathcal{G}_S\right)\right| \to 0 \text{ and } \max_{\mathcal{G}_S \in \mathcal{C}} \left|\widehat{\psi}_3\left(\mathcal{G}_S\right) - \psi_3\left(\mathcal{G}_S\right)\right| \to 0$$

almost uniformly. So for every $\varepsilon > 0$, there is a measurable set $A \subset \Omega$ with $\mathbb{P}(A) \geq 1 - \varepsilon$ such that for all large n,

$$\max_{l \in \{1,2,3\}} \max_{\mathcal{G}_S \in \mathcal{C}} \left|\widehat{\psi}_l\left(\mathcal{G}_S\right) - \psi_l\left(\mathcal{G}_S\right)\right| \leq \frac{\delta}{2}$$

uniformly on A. Thus, it follows that for every $l \in \{1, 2, 3\}$,

$$\begin{aligned}&\lim_{n\to\infty} \mathbb{P}\left(\max_{\mathcal{G}_S \in \widehat{\mathcal{C}_2}\setminus\mathcal{C}_2} \psi_l\left(\mathcal{G}_S\right) > \delta, \max_{\mathcal{G}_S \in \widehat{\mathcal{C}_2}\setminus\mathcal{C}_2} \sqrt{T_n}\widehat{\psi}_l\left(\mathcal{G}_S\right) \leq t_n\right) \\ \leq & \lim_{n\to\infty} \mathbb{P}\left(\left\{\max_{\mathcal{G}_S \in \widehat{\mathcal{C}_2}\setminus\mathcal{C}_2} \psi_l\left(\mathcal{G}_S\right) > \delta, \max_{\mathcal{G}_S \in \widehat{\mathcal{C}_2}\setminus\mathcal{C}_2} \sqrt{T_n}\widehat{\psi}_l\left(\mathcal{G}_S\right) \leq t_n\right\} \cap A\right) + \mathbb{P}(A^c) \\ \leq & \lim_{n\to\infty} \mathbb{P}\left(\frac{\delta}{2} \leq \max_{\mathcal{G}_S \in \widehat{\mathcal{C}_2}\setminus\mathcal{C}_2} \widehat{\psi}_l\left(\mathcal{G}_S\right) \leq \frac{t_n}{\sqrt{T_n}}\right) + \varepsilon = \varepsilon.\end{aligned}$$

Since ε can be arbitrarily small, we have that

$$\mathbb{P}\left(\max_{\mathcal{G}_S \in \widehat{\mathcal{C}_2} \setminus \mathcal{C}_2} \psi_l(\mathcal{G}_S) > \delta, \max_{\mathcal{G}_S \in \widehat{\mathcal{C}_2} \setminus \mathcal{C}_2} \sqrt{T_n}\widehat{\psi}_l(\mathcal{G}_S) \leq t_n\right) \to 0.$$

This implies $\mathbb{P}(\widehat{\mathcal{C}_2} \setminus \mathcal{C}_2 \neq \varnothing) \to 0$. Thus,

$$\mathbb{P}\left(\widehat{\mathcal{C}_2} \neq \mathcal{C}_2\right) \leq \mathbb{P}\left(\widehat{\mathcal{C}_2} \setminus \mathcal{C}_2 \neq \varnothing\right) + \mathbb{P}\left(\mathcal{C}_2 \setminus \widehat{\mathcal{C}_2} \neq \varnothing\right) \to 0.$$

■

B.5 Partially Valid Instruments for Multivalued Ordered Treatments

Here we extend the analysis in Section 3.3 to multivalued ordered treatments. We follow the setup in Section A.1. Consider the following generalized version of Definition 3.2.

Definition B.2 *Suppose the instrument Z is pairwise valid for the (multivalued ordered) treatment D with the largest validity pair set $\mathscr{Z}_{\bar{M}}$. If there is a validity pair set*

$$\mathscr{Z}_M = \{(z_{k_1}, z_{k_2}), (z_{k_2}, z_{k_3}), \ldots, (z_{k_{M-1}}, z_{k_M})\}$$

*for some $M > 0$, then the instrument Z is called a **partially valid instrument** for the treatment D. The set $\mathcal{Z}_M = \{z_{k_1}, \ldots, z_{k_M}\}$ is called a **validity value set** of Z.*

Assumption B.1 *The validity value set $\mathcal{Z}_M$ satisfies that*

$$E[g(Z_i)D_i | Z_i \in \mathcal{Z}_M] - E[D_i | Z_i \in \mathcal{Z}_M] \cdot E[g(Z_i) | Z_i \in \mathcal{Z}_M] \neq 0. \tag{B.24}$$

Suppose that we have access to a consistent estimator $\widehat{\mathcal{Z}_0}$ of the validity value set $\mathcal{Z}_M$, that is, $\mathbb{P}(\widehat{\mathcal{Z}_0} = \mathcal{Z}_M) \to 1$. Then we can use $\widehat{\mathcal{Z}_0}$ to construct a VSIV estimator, $\widehat{\theta}_1$, for a weighted average of ACRs based on model (3.14), where D is now a multivalued ordered treatment. The following theorem presents the asymptotic properties of the VSIV estimator, generalizing Theorem 3.3. Theorem B.1 is an extension of Theorem 2 of Imbens and Angrist (1994) and Theorem 2 of Angrist and Imbens (1995) to the case where the instrument is partially but not fully valid.

Theorem B.1 *Suppose that the instrument Z is partially valid for the treatment D as defined in Definition B.2 with a validity value set $\mathcal{Z}_M = \{z_{k_1}, \ldots, z_{k_M}\}$, and that the estimator $\widehat{\mathcal{Z}_0}$*

for $\mathcal{Z}_M$ satisfies $\mathbb{P}(\widehat{\mathcal{Z}_0} = \mathcal{Z}_M) \to 1$. Under Assumptions A.2 and B.1, it follows that $\widehat{\theta}_1 \xrightarrow{p} \theta_1$, where

$$\theta_1 = \frac{E\left[g\left(Z_i\right) Y_i | Z_i \in \mathcal{Z}_M\right] - E\left[Y_i | Z_i \in \mathcal{Z}_M\right] E\left[g\left(Z_i\right) | Z_i \in \mathcal{Z}_M\right]}{E\left[g\left(Z_i\right) D_i | Z_i \in \mathcal{Z}_M\right] - E\left[D_i | Z_i \in \mathcal{Z}_M\right] E\left[g\left(Z_i\right) | Z_i \in \mathcal{Z}_M\right]}.$$

Also, $\sqrt{n}(\widehat{\theta}_1 - \theta_1) \xrightarrow{d} N\left(0, \Sigma_1\right)$, where Σ_1 is provided in (B.25). In addition, the quantity θ_1 can be interpreted as the weighted average of $\{\beta_{k_2,k_1}, \ldots, \beta_{k_M,k_{M-1}}\}$ defined in (A.1). Specifically, $\theta_1 = \sum_{m=1}^{M-1} \mu_m \beta_{k_{m+1},k_m}$ with

$$\mu_m = \frac{\left[p\left(z_{k_{m+1}}\right) - p\left(z_{k_m}\right)\right] \sum_{l=m}^{M-1} \mathbb{P}\left(Z_i = z_{k_{l+1}} | Z_i \in \mathcal{Z}_M\right) \left\{g\left(z_{k_{l+1}}\right) - E\left[g\left(Z_i\right) | Z_i \in \mathcal{Z}_M\right]\right\}}{\sum_{l=1}^{M} \mathbb{P}\left(Z_i = z_{k_l} | Z_i \in \mathcal{Z}_M\right) p\left(z_{k_l}\right) \left\{g\left(z_{k_l}\right) - E\left[g\left(Z_i\right) | Z_i \in \mathcal{Z}_M\right]\right\}},$$

$p\left(z_k\right) = E\left[D_i | Z_i = z_k\right]$, and $\sum_{m=1}^{M-1} \mu_m = 1$.

Proof of Theorem B.1. By the formula of the VSIV estimator in (3.15),

$$\widehat{\theta}_1 = \frac{\frac{n_z}{n} \frac{1}{n} \sum_{i=1}^{n} g\left(Z_i\right) Y_i 1\left\{Z_i \in \widehat{\mathcal{Z}_0}\right\} - \bar{Y}_{\widehat{\mathcal{Z}_0}} \frac{1}{n} \sum_{i=1}^{n} g\left(Z_i\right) 1\left\{Z_i \in \widehat{\mathcal{Z}_0}\right\}}{\frac{n_z}{n} \frac{1}{n} \sum_{i=1}^{n} g\left(Z_i\right) D_i 1\left\{Z_i \in \widehat{\mathcal{Z}_0}\right\} - \bar{D}_{\widehat{\mathcal{Z}_0}} \frac{1}{n} \sum_{i=1}^{n} g\left(Z_i\right) 1\left\{Z_i \in \widehat{\mathcal{Z}_0}\right\}},$$

where

$$\bar{Y}_{\widehat{\mathcal{Z}_0}} = \frac{1}{n} \sum_{i=1}^{n} Y_i 1\left\{Z_i \in \widehat{\mathcal{Z}_0}\right\} \text{ and } \bar{D}_{\widehat{\mathcal{Z}_0}} = \frac{1}{n} \sum_{i=1}^{n} D_i 1\left\{Z_i \in \widehat{\mathcal{Z}_0}\right\}.$$

We first have

$$\begin{aligned}
&\frac{1}{n} \sum_{i=1}^{n} g\left(Z_i\right) Y_i 1\left\{Z_i \in \widehat{\mathcal{Z}_0}\right\} \\
&= \frac{1}{n} \sum_{i=1}^{n} g\left(Z_i\right) Y_i 1\left\{Z_i \in \mathcal{Z}_M\right\} + \left[\frac{1}{n} \sum_{i=1}^{n} g\left(Z_i\right) Y_i \left\{1\left\{Z_i \in \widehat{\mathcal{Z}_0}\right\} - 1\left\{Z_i \in \mathcal{Z}_M\right\}\right\}\right]
\end{aligned}$$

with

$$\left|\frac{1}{n} \sum_{i=1}^{n} g\left(Z_i\right) Y_i \left\{1\left\{Z_i \in \widehat{\mathcal{Z}_0}\right\} - 1\left\{Z_i \in \mathcal{Z}_M\right\}\right\}\right| \leq \frac{1}{n} \sum_{i=1}^{n} \left|g\left(Z_i\right) Y_i\right| 1\left\{\widehat{\mathcal{Z}_0} \neq \mathcal{Z}_M\right\}.$$

Since $n^{-1} \sum_{i=1}^{n} \left|g\left(Z_i\right) Y_i\right| \xrightarrow{p} E\left[\left|g\left(Z_i\right) Y_i\right|\right]$ and for every small $\varepsilon > 0$,

$$\mathbb{P}\left(1\left\{\widehat{\mathcal{Z}_0} \neq \mathcal{Z}_M\right\} > \varepsilon\right) = \mathbb{P}\left(\widehat{\mathcal{Z}_0} \neq \mathcal{Z}_M\right) \to 0,$$

we have that

$$\frac{1}{n}\sum_{i=1}^{n} g\left(Z_i\right) Y_i 1\left\{Z_i \in \widehat{\mathcal{Z}}_0\right\} = \frac{1}{n}\sum_{i=1}^{n} g\left(Z_i\right) Y_i 1\left\{Z_i \in \mathcal{Z}_M\right\} + o_p\left(1\right)$$
$$\xrightarrow{p} E\left[g\left(Z_i\right) Y_i 1\left\{Z_i \in \mathcal{Z}_M\right\}\right].$$

Recall that $n_z = \sum_{i=1}^{n} 1\{Z_i \in \widehat{\mathcal{Z}}_0\}$. Then we can show that $n_z/n \xrightarrow{p} \mathbb{P}(Z_i \in \mathcal{Z}_M)$ as $n \to \infty$. Similarly, we have that $\bar{Y}_{\widehat{\mathcal{Z}}_0} \xrightarrow{p} E\left[Y_i 1\left\{Z_i \in \mathcal{Z}_M\right\}\right]$, $\bar{D}_{\widehat{\mathcal{Z}}_0} \xrightarrow{p} E\left[D_i 1\left\{Z_i \in \mathcal{Z}_M\right\}\right]$, $n^{-1}\sum_{i=1}^{n} g\left(Z_i\right) 1\{Z_i \in \widehat{\mathcal{Z}}_0\} \xrightarrow{p} E\left[g\left(Z_i\right) 1\left\{Z_i \in \mathcal{Z}_M\right\}\right]$, and $n^{-1}\sum_{i=1}^{n} g\left(Z_i\right) D_i 1\{Z_i \in \widehat{\mathcal{Z}}_0\} \xrightarrow{p} E\left[g\left(Z_i\right) D_i 1\left\{Z_i \in \mathcal{Z}_M\right\}\right]$. Thus, it follows that

$$\widehat{\theta}_1 \xrightarrow{p} \frac{\frac{E[g(Z_i)Y_i 1\{Z_i\in\mathcal{Z}_M\}]}{\mathbb{P}(Z_i\in\mathcal{Z}_M)} - \frac{E[Y_i 1\{Z_i\in\mathcal{Z}_M\}]}{\mathbb{P}(Z_i\in\mathcal{Z}_M)}\frac{E[g(Z_i)1\{Z_i\in\mathcal{Z}_M\}]}{\mathbb{P}(Z_i\in\mathcal{Z}_M)}}{\frac{E[g(Z_i)D_i 1\{Z_i\in\mathcal{Z}_M\}]}{\mathbb{P}(Z_i\in\mathcal{Z}_M)} - \frac{E[D_i 1\{Z_i\in\mathcal{Z}_M\}]}{\mathbb{P}(Z_i\in\mathcal{Z}_M)}\frac{E[g(Z_i)1\{Z_i\in\mathcal{Z}_M\}]}{\mathbb{P}(Z_i\in\mathcal{Z}_M)}} = \theta_1.$$

Next, we derive the asymptotic distribution of $\sqrt{n}(\widehat{\theta}_1 - \theta_1)$. Define a function $f : \mathbb{R}^6 \to \mathbb{R}$ by

$$f\left(x\right) = \frac{x_1/x_6 - x_2 x_3/x_6^2}{x_4/x_6 - x_5 x_3/x_6^2}$$

for every $x \in \mathbb{R}^6$ with $x = \left(x_1, x_2, x_3, x_4, x_5, x_6\right)^T$ such that $f(x)$ is well defined. We can obtain the gradient of f, denoted f', by $f'\left(x\right) = \left(f_1'\left(x\right), f_2'\left(x\right), f_3'\left(x\right), f_4'\left(x\right), f_5'\left(x\right), f_6'\left(x\right)\right)^T$, where

$$f_1'\left(x\right) = \frac{x_6}{x_4 x_6 - x_5 x_3}, f_2'\left(x\right) = \frac{-x_3}{x_4 x_6 - x_5 x_3}, f_3'\left(x\right) = \frac{-x_2 x_4 x_6 + x_5 x_1 x_6}{\left(x_4 x_6 - x_5 x_3\right)^2},$$
$$f_4'\left(x\right) = -\frac{\left(x_1 x_6 - x_2 x_3\right) x_6}{\left(x_4 x_6 - x_5 x_3\right)^2}, f_5'\left(x\right) = \frac{x_3\left(x_1 x_6 - x_2 x_3\right)}{\left(x_4 x_6 - x_5 x_3\right)^2}, \text{ and } f_6'\left(x\right) = \frac{-x_1 x_5 x_3 + x_2 x_3 x_4}{\left(x_4 x_6 - x_5 x_3\right)^2}$$

for every $x = \left(x_1, x_2, x_3, x_4, x_5, x_6\right)^T$ such that all the above derivatives are well defined. Then we can rewrite

$$\sqrt{n}(\widehat{\theta}_1 - \theta_1) = \sqrt{n}\left\{f\left(\widehat{W}_n\right) - f\left(W\right)\right\},$$

where

$$\widehat{W}_n = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n g(Z_i) Y_i 1\left\{Z_i \in \widehat{\mathcal{Z}}_0\right\} \\ \bar{Y}_{\widehat{\mathcal{Z}}_0} \\ \frac{1}{n}\sum_{i=1}^n g(Z_i) 1\left\{Z_i \in \widehat{\mathcal{Z}}_0\right\} \\ \frac{1}{n}\sum_{i=1}^n g(Z_i) D_i 1\left\{Z_i \in \widehat{\mathcal{Z}}_0\right\} \\ \bar{D}_{\widehat{\mathcal{Z}}_0} \\ \frac{1}{n}\sum_{i=1}^n 1\left\{Z_i \in \widehat{\mathcal{Z}}_0\right\} \end{pmatrix} \text{ and } W = \begin{pmatrix} E\left[g(Z_i) Y_i 1\left\{Z_i \in \mathcal{Z}_M\right\}\right] \\ E\left[Y_i 1\left\{Z_i \in \mathcal{Z}_M\right\}\right] \\ E\left[g(Z_i) 1\left\{Z_i \in \mathcal{Z}_M\right\}\right] \\ E\left[g(Z_i) D_i 1\left\{Z_i \in \mathcal{Z}_M\right\}\right] \\ E\left[D_i 1\left\{Z_i \in \mathcal{Z}_M\right\}\right] \\ E\left[1\left\{Z_i \in \mathcal{Z}_M\right\}\right] \end{pmatrix}.$$

For every small $\varepsilon > 0$, we have $\mathbb{P}(\sqrt{n}1\{\widehat{\mathcal{Z}_0} \neq \mathcal{Z}_M\} > \varepsilon) = \mathbb{P}(\widehat{\mathcal{Z}_0} \neq \mathcal{Z}_M) \to 0$. With $n^{-1}\sum_{i=1}^n |g(Z_i) Y_i| \xrightarrow{p} E[|g(Z_i) Y_i|]$, we have that

$$\begin{aligned} &\sqrt{n}\left|\frac{1}{n}\sum_{i=1}^n g(Z_i) Y_i 1\left\{Z_i \in \widehat{\mathcal{Z}_0}\right\} - \frac{1}{n}\sum_{i=1}^n g(Z_i) Y_i 1\left\{Z_i \in \mathcal{Z}_M\right\}\right| \\ =&\sqrt{n}\left|\frac{1}{n}\sum_{i=1}^n g(Z_i) Y_i\left[1\left\{Z_i \in \widehat{\mathcal{Z}_0}\right\} - 1\left\{Z_i \in \mathcal{Z}_M\right\}\right]\right| \\ \leq&\frac{1}{n}\sum_{i=1}^n |g(Z_i) Y_i|\left(\sqrt{n}1\left\{\widehat{\mathcal{Z}_0} \neq \mathcal{Z}_M\right\}\right) = o_p(1). \end{aligned}$$

Similarly, we have that

$$\begin{aligned} &\sqrt{n}\left(\widehat{W}_n - W\right) \\ =&\sqrt{n}\frac{1}{n}\sum_{i=1}^n \begin{pmatrix} g(Z_i) Y_i 1\left\{Z_i \in \mathcal{Z}_M\right\} - E\left[g(Z_i) Y_i 1\left\{Z_i \in \mathcal{Z}_M\right\}\right] \\ Y_i 1\left\{Z_i \in \mathcal{Z}_M\right\} - E\left[Y_i 1\left\{Z_i \in \mathcal{Z}_M\right\}\right] \\ g(Z_i) 1\left\{Z_i \in \mathcal{Z}_M\right\} - E\left[g(Z_i) 1\left\{Z_i \in \mathcal{Z}_M\right\}\right] \\ g(Z_i) D_i 1\left\{Z_i \in \mathcal{Z}_M\right\} - E\left[g(Z_i) D_i 1\left\{Z_i \in \mathcal{Z}_M\right\}\right] \\ D_i 1\left\{Z_i \in \mathcal{Z}_M\right\} - E\left[D_i 1\left\{Z_i \in \mathcal{Z}_M\right\}\right] \\ 1\left\{Z_i \in \mathcal{Z}_M\right\} - E\left[1\left\{Z_i \in \mathcal{Z}_M\right\}\right] \end{pmatrix} + o_p(1) \xrightarrow{d} N(0, \Sigma), \end{aligned}$$

where $\Sigma = E\left[VV^T\right]$ and

$$V = \begin{pmatrix} g(Z_i) Y_i 1\{Z_i \in \mathcal{Z}_M\} - E[g(Z_i) Y_i 1\{Z_i \in \mathcal{Z}_M\}] \\ Y_i 1\{Z_i \in \mathcal{Z}_M\} - E[Y_i 1\{Z_i \in \mathcal{Z}_M\}] \\ g(Z_i) 1\{Z_i \in \mathcal{Z}_M\} - E[g(Z_i) 1\{Z_i \in \mathcal{Z}_M\}] \\ g(Z_i) D_i 1\{Z_i \in \mathcal{Z}_M\} - E[g(Z_i) D_i 1\{Z_i \in \mathcal{Z}_M\}] \\ D_i 1\{Z_i \in \mathcal{Z}_M\} - E[D_i 1\{Z_i \in \mathcal{Z}_M\}] \\ 1\{Z_i \in \mathcal{Z}_M\} - E[1\{Z_i \in \mathcal{Z}_M\}] \end{pmatrix}.$$

By multivariate delta method, we have that

$$\sqrt{n}(\widehat{\theta}_1 - \theta_1) = \sqrt{n}\left\{f\left(\widehat{W}_n\right) - f(W)\right\} \xrightarrow{d} f'(W)^T \cdot N(0, \Sigma). \tag{B.25}$$

Now we follow the strategy of Imbens and Angrist (1994) and have that

$$\begin{aligned} &\frac{E[g(Z_i) Y_i 1\{Z_i \in \mathcal{Z}_M\}]}{\mathbb{P}(Z_i \in \mathcal{Z}_M)} - \frac{E[Y_i 1\{Z_i \in \mathcal{Z}_M\}]}{\mathbb{P}(Z_i \in \mathcal{Z}_M)} \frac{E[g(Z_i) 1\{Z_i \in \mathcal{Z}_M\}]}{\mathbb{P}(Z_i \in \mathcal{Z}_M)} \\ =&\frac{\sum_{k=1}^K \mathbb{P}(Z_i = z_k) E[Y_i 1\{Z_i \in \mathcal{Z}_M\} | Z_i = z_k] \left\{g(z_k) 1\{z_k \in \mathcal{Z}_M\} - \frac{E[g(Z_i) 1\{Z_i \in \mathcal{Z}_M\}]}{\mathbb{P}(Z_i \in \mathcal{Z}_M)}\right\}}{\mathbb{P}(Z_i \in \mathcal{Z}_M)} \\ =&\sum_{m=1}^M \mathbb{P}(Z_i = z_{k_m} | Z_i \in \mathcal{Z}_M) E[Y_i | Z_i = z_{k_m}] \{g(z_{k_m}) - E[g(Z_i) | Z_i \in \mathcal{Z}_M]\}. \end{aligned}$$

Then we write

$$\begin{aligned} &\sum_{m=1}^M \mathbb{P}(Z_i = z_{k_m} | Z_i \in \mathcal{Z}_M) E[Y_i | Z_i = z_{k_m}] \{g(z_{k_m}) - E[g(Z_i) | Z_i \in \mathcal{Z}_M]\} \\ =&\sum_{m=1}^{M-1} \mathbb{P}\left(Z_i = z_{k_{m+1}} | Z_i \in \mathcal{Z}_M\right) E\left[Y_i | Z_i = z_{k_{m+1}}\right] \left\{g\left(z_{k_{m+1}}\right) - E[g(Z_i) | Z_i \in \mathcal{Z}_M]\right\} \\ &+ \mathbb{P}(Z_i = z_{k_1} | Z_i \in \mathcal{Z}_M) E[Y_i | Z_i = z_{k_1}] \{g(z_{k_1}) - E[g(Z_i) | Z_i \in \mathcal{Z}_M]\}. \end{aligned} \tag{B.26}$$

By (A.1), we have

$$\begin{aligned} E\left[Y_i | Z_i = z_{k_{m+1}}\right] &= \beta_{k_{m+1},k_m}\left(E\left[D_i | Z_i = z_{k_{m+1}}\right] - E[D_i | Z_i = z_{k_m}]\right) + E[Y_i | Z_i = z_{k_m}] \\ &= \sum_{l=1}^m \beta_{k_{l+1},k_l}\left(E\left[D_i | Z_i = z_{k_{l+1}}\right] - E[D_i | Z_i = z_{k_l}]\right) + E[Y_i | Z_i = z_{k_1}], \end{aligned}$$

and thus it follows that

$$\begin{aligned}
&\sum_{m=1}^{M-1} \mathbb{P}\left(Z_i = z_{k_{m+1}} | Z_i \in \mathcal{Z}_M\right) E\left[Y_i | Z_i = z_{k_{m+1}}\right] \left\{g\left(z_{k_{m+1}}\right) - E\left[g\left(Z_i\right) | Z_i \in \mathcal{Z}_M\right]\right\} \\
&= \sum_{m=1}^{M-1} \left\{ \mathbb{P}\left(Z_i = z_{k_{m+1}} | Z_i \in \mathcal{Z}_M\right) \left\{ \sum_{l=1}^{m} \beta_{k_{l+1},k_l} \left[p\left(z_{k_{l+1}}\right) - p\left(z_{k_l}\right)\right] \right\} \right. \\
&\quad \left. \cdot \left\{g\left(z_{k_{m+1}}\right) - E\left[g\left(Z_i\right) | Z_i \in \mathcal{Z}_M\right]\right\} \right\} \\
&\quad + \sum_{m=1}^{M-1} \mathbb{P}\left(Z_i = z_{k_{m+1}} | Z_i \in \mathcal{Z}_M\right) E[Y_i | Z_i = z_{k_1}] \left\{g\left(z_{k_{m+1}}\right) - E\left[g\left(Z_i\right) | Z_i \in \mathcal{Z}_M\right]\right\}.
\end{aligned}$$

By (B.26), this implies that

$$\begin{aligned}
&\sum_{m=1}^{M} \mathbb{P}\left(Z_i = z_{k_m} | Z_i \in \mathcal{Z}_M\right) E\left[Y_i | Z_i = z_{k_m}\right] \left\{g\left(z_{k_m}\right) - E\left[g\left(Z_i\right) | Z_i \in \mathcal{Z}_M\right]\right\} \\
&= \sum_{m=1}^{M-1} \left\{ \mathbb{P}\left(Z_i = z_{k_{m+1}} | Z_i \in \mathcal{Z}_M\right) \left\{ \sum_{l=1}^{m} \beta_{k_{l+1},k_l} \left[p\left(z_{k_{l+1}}\right) - p\left(z_{k_l}\right)\right] \right\} \right. \\
&\quad \left. \cdot \left\{g\left(z_{k_{m+1}}\right) - E\left[g\left(Z_i\right) | Z_i \in \mathcal{Z}_M\right]\right\} \right\}, \qquad \text{(B.27)}
\end{aligned}$$

where we use $\sum_{m=1}^{M} \mathbb{P}\left(Z_i = z_{k_m} | Z_i \in \mathcal{Z}_M\right) \left\{g\left(z_{k_m}\right) - E\left[g\left(Z_i\right) | Z_i \in \mathcal{Z}_M\right]\right\} = 0$. By rewriting (B.27), we obtain

$$\begin{aligned}
&\sum_{m=1}^{M-1} \mathbb{P}\left(Z_i = z_{k_{m+1}} | Z_i \in \mathcal{Z}_M\right) \left\{ \sum_{l=1}^{m} \beta_{k_{l+1},k_l} \left[p\left(z_{k_{l+1}}\right) - p\left(z_{k_l}\right)\right] \right\} \tilde{g}\left(z_{k_{m+1}}\right) \\
&= \mathbb{P}\left(Z_i = z_{k_2} | Z_i \in \mathcal{Z}_M\right) \left\{\beta_{k_2,k_1} \left[p\left(z_{k_2}\right) - p\left(z_{k_1}\right)\right]\right\} \tilde{g}\left(z_{k_2}\right) + \cdots \\
&\quad + \mathbb{P}\left(Z_i = z_{k_M} | Z_i \in \mathcal{Z}_M\right) \left\{ \sum_{l=1}^{M-1} \beta_{k_{l+1},k_l} \left[p\left(z_{k_{l+1}}\right) - p\left(z_{k_l}\right)\right] \right\} \tilde{g}\left(z_{k_M}\right) \\
&= \sum_{m=1}^{M-1} \left\{ \beta_{k_{m+1},k_m} \left[p\left(z_{k_{m+1}}\right) - p\left(z_{k_m}\right)\right] \sum_{l=m}^{M-1} \mathbb{P}\left(Z_i = z_{k_{l+1}} | Z_i \in \mathcal{Z}_M\right) \tilde{g}\left(z_{k_{l+1}}\right) \right\},
\end{aligned}$$

where $\tilde{g}\left(z\right) = g\left(z\right) - E\left[g\left(Z_i\right) | Z_i \in \mathcal{Z}_M\right]$ for all z. Similarly, we have

$$\begin{aligned}
&\frac{E\left[g\left(Z_i\right) D_i 1\left\{Z_i \in \mathcal{Z}_M\right\}\right]}{\mathbb{P}\left(Z_i \in \mathcal{Z}_M\right)} - \frac{E\left[D_i 1\left\{Z_i \in \mathcal{Z}_M\right\}\right]}{\mathbb{P}\left(Z_i \in \mathcal{Z}_M\right)} \frac{E\left[g\left(Z_i\right) 1\left\{Z_i \in \mathcal{Z}_M\right\}\right]}{\mathbb{P}\left(Z_i \in \mathcal{Z}_M\right)} \\
&= \sum_{m=1}^{M} \mathbb{P}\left(Z_i = z_{k_m} | Z_i \in \mathcal{Z}_M\right) p\left(z_{k_m}\right) \left\{g\left(z_{k_m}\right) - E\left[g\left(Z_i\right) | Z_i \in \mathcal{Z}_M\right]\right\},
\end{aligned}$$

which is nonzero by Assumption B.1. Thus, we have $\theta_1 = \sum_{m=1}^{M-1} \mu_m \beta_{k_{m+1},k_m}$ with

$$\mu_m = \frac{\left[p\left(z_{k_{m+1}}\right) - p\left(z_{k_m}\right)\right] \sum_{l=m}^{M-1} \mathbb{P}\left(Z_i = z_{k_{l+1}} | Z_i \in \mathcal{Z}_M\right) \left\{g\left(z_{k_{l+1}}\right) - E\left[g\left(Z_i\right) | Z_i \in \mathcal{Z}_M\right]\right\}}{\sum_{l=1}^{M} \mathbb{P}\left(Z_i = z_{k_l} | Z_i \in \mathcal{Z}_M\right) p\left(z_{k_l}\right) \left\{g\left(z_{k_l}\right) - E\left[g\left(Z_i\right) | Z_i \in \mathcal{Z}_M\right]\right\}}.$$

Now we show that $\sum_{m=1}^{M-1} \mu_m = 1$. First, we have that

$$\begin{aligned}
&\sum_{m=1}^{M-1} \left[p\left(z_{k_{m+1}}\right) - p\left(z_{k_m}\right)\right] \sum_{l=m}^{M-1} \mathbb{P}\left(Z_i = z_{k_{l+1}} | Z_i \in \mathcal{Z}_M\right) \left\{g\left(z_{k_{l+1}}\right) - E\left[g\left(Z_i\right) | Z_i \in \mathcal{Z}_M\right]\right\} \\
&= \left[p\left(z_{k_2}\right) - p\left(z_{k_1}\right)\right] \sum_{l=1}^{M-1} \mathbb{P}\left(Z_i = z_{k_{l+1}} | Z_i \in \mathcal{Z}_M\right) \left\{g\left(z_{k_{l+1}}\right) - E\left[g\left(Z_i\right) | Z_i \in \mathcal{Z}_M\right]\right\} + \cdots \\
&\quad + \left[p\left(z_{k_M}\right) - p\left(z_{k_{M-1}}\right)\right] \mathbb{P}\left(Z_i = z_{k_M} | Z_i \in \mathcal{Z}_M\right) \left\{g\left(z_{k_M}\right) - E\left[g\left(Z_i\right) | Z_i \in \mathcal{Z}_M\right]\right\} \\
&= \sum_{l=2}^{M} \mathbb{P}\left(Z_i = z_{k_l} | Z_i \in \mathcal{Z}_M\right) p\left(z_{k_l}\right) \left\{g\left(z_{k_l}\right) - E\left[g\left(Z_i\right) | Z_i \in \mathcal{Z}_M\right]\right\} \\
&\quad - p\left(z_{k_1}\right) \sum_{l=2}^{M} \mathbb{P}\left(Z_i = z_{k_l} | Z_i \in \mathcal{Z}_M\right) \left\{g\left(z_{k_l}\right) - E\left[g\left(Z_i\right) | Z_i \in \mathcal{Z}_M\right]\right\} \\
&= \sum_{l=1}^{M} \mathbb{P}\left(Z_i = z_{k_l} | Z_i \in \mathcal{Z}_M\right) p\left(z_{k_l}\right) \left\{g\left(z_{k_l}\right) - E\left[g\left(Z_i\right) | Z_i \in \mathcal{Z}_M\right]\right\},
\end{aligned}$$

where we use the equality that $\sum_{l=1}^{M} \mathbb{P}\left(Z_i = z_{k_l} | Z_i \in \mathcal{Z}_M\right) \left\{g\left(z_{k_l}\right) - E\left[g\left(Z_i\right) | Z_i \in \mathcal{Z}_M\right]\right\} = 0$. This implies that $\sum_{m=1}^{M-1} \mu_m = 1$. ■

C Proofs and Supplementary Results for Appendix A.2

C.1 Proofs for Appendix A.2

Proof of Lemma A.2. (i) $\Leftrightarrow$ (ii). We closely follow the proof for "(i) $\Leftrightarrow$ (ii)" in Theorem T-3 of Heckman and Pinto (2018). By Lemma L-5 of Heckman and Pinto (2018), if $B_{d(k,k')}$ is lonesum, then no 2×2 sub-matrix of $B_{d(k,k')}$ takes the form

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \text{ or } \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \tag{C.1}$$

Since $B_{d(k,k')} = 1\{\mathcal{K}_{(k,k')}R = d\}$, (i) $\Rightarrow$ (ii). Suppose (ii) holds. Then no 2×2 sub-matrix of $B_{d(k,k')}$ takes the form in (C.1) by the definition of $B_{d(k,k')}$. By Lemmas L-6 and L-8 of Heckman and Pinto (2018), (i) holds.

(i) $\Rightarrow$ (iii) $\Rightarrow$ (ii). If for every $d \in \mathcal{D}$, $B_{d(k,k')}$ is lonesum, by Lemma L-9 of Heckman and Pinto (2018),

$$B_{d(k,k')}(1,l) \leq B_{d(k,k')}(2,l) \text{ for all } l, \text{ or } B_{d(k,k')}(1,l) \geq B_{d(k,k')}(2,l) \text{ for all } l.$$

Because the value of $(D_{z_k}, D_{z_{k'}})$ must be equal to $(\mathcal{K}_{(k,k')}R(1,l), \mathcal{K}_{(k,k')}R(2,l))$ for some l, it follows that

$$1\{D_{z_k} = d\} \leq 1\left\{D_{z_{k'}} = d\right\} \text{ or } 1\{D_{z_k} = d\} \geq 1\left\{D_{z_{k'}} = d\right\}.$$

Thus the following sub-matrices will not occur in $\mathcal{K}_{(k,k')}R$:

$$\begin{pmatrix} d & d' \\ d'' & d \end{pmatrix} \text{ or } \begin{pmatrix} d' & d \\ d & d'' \end{pmatrix},$$

where $d' \neq d$ and $d'' \neq d$. ■

Proof of Theorem A.3. The proof follows a strategy similar to that of the proof of Theorem T-6 in Heckman and Pinto (2018). We first write

$$\mathbb{P}\left(\mathcal{M}_{(k,k')}S \in \Sigma_{d(k,k')}(t)\right) = b_{d(k,k')}(t) P_{S(k,k')}. \tag{C.2}$$

Also, since

$$\begin{aligned} &E\left[\kappa\left(Y_d(z_k, z_{k'})\right) 1\left\{\mathcal{M}_{(k,k')}S \in \Sigma_{d(k,k')}(t)\right\}\right] \\ &= E\left[E\left[\kappa\left(Y_d(z_k, z_{k'})\right) 1\left\{\mathcal{M}_{(k,k')}S \in \Sigma_{d(k,k')}(t)\right\} | 1\left\{\mathcal{M}_{(k,k')}S \in \Sigma_{d(k,k')}(t)\right\}\right]\right] \\ &= E\left[\kappa\left(Y_d(z_k, z_{k'})\right) | \mathcal{M}_{(k,k')}S \in \Sigma_{d(k,k')}(t)\right] \cdot \mathbb{P}\left(\mathcal{M}_{(k,k')}S \in \Sigma_{d(k,k')}(t)\right) \end{aligned}$$

and

$$\begin{aligned} &E\left[\kappa\left(Y_d(z_k, z_{k'})\right) 1\left\{\mathcal{M}_{(k,k')}S \in \Sigma_{d(k,k')}(t)\right\}\right] \\ &= E\left[\kappa\left(Y_d(z_k, z_{k'})\right) \sum_{l=1}^{L_{(k,k')}} 1\left\{\mathcal{M}_{(k,k')}S = s_l\right\} 1\left\{s_l \in \Sigma_{d(k,k')}(t)\right\}\right] = b_{d(k,k')}(t) Q_{S(k,k')}(d), \end{aligned}$$

we have that

$$E\left[\kappa\left(Y_d(z_k, z_{k'})\right) | \mathcal{M}_{(k,k')} S \in \Sigma_{d(k,k')}(t)\right] = \frac{b_{d(k,k')}(t) Q_{S(k,k')}(d)}{b_{d(k,k')}(t) P_{S(k,k')}}. \tag{C.3}$$

Now we suppose $(z_k, z_{k'}) \in \mathscr{Z}_{\bar{M}}$. By definition, we have $P_{Z(k,k')}(d) = B_{d(k,k')} P_{S(k,k')}$ and $Q_{Z(k,k')}(d) = B_{d(k,k')} Q_{S(k,k')}(d)$, so by Lemma L-2 of Heckman and Pinto (2018),

$$b_{d(k,k')}(t) P_{S(k,k')} = b_{d(k,k')}(t) \left[B^{+}_{d(k,k')} P_{Z(k,k')}(d) + \left(I - B^{+}_{d(k,k')} B_{d(k,k')}\right) \lambda_P\right] \text{ and}$$
$$b_{d(k,k')}(t) Q_{S(k,k')}(d) = b_{d(k,k')}(t) \left[B^{+}_{d(k,k')} Q_{Z(k,k')}(d) + \left(I - B^{+}_{d(k,k')} B_{d(k,k')}\right) \lambda_Q\right],$$

where λ_P and λ_Q are some real-valued vectors.

We next show that $b_{d(k,k')}(t)\left[I - B^{+}_{d(k,k')} B_{d(k,k')}\right] = 0$. First, by the proof of Lemma L-16 of Heckman and Pinto (2018) and Lemma A.2 in this paper, if $B_{d(k,k')}(\cdot, l)$ and $B_{d(k,k')}(\cdot, l')$ have the same sum, then these two vectors are identical. Thus, by the definition of the set $\Sigma_{d(k,k')}(t)$, for all $s_l, s_{l'} \in \Sigma_{d(k,k')}(t)$, $B_{d(k,k')}(\cdot, l) = B_{d(k,k')}(\cdot, l')$. Let $C_{d(k,k')}(t) = B_{d(k,k')}(\cdot, l)$ with l satisfying that $s_l \in \Sigma_{d(k,k')}(t)$, where s_l is the lth column of $\mathcal{K}_{(k,k')} R$. Let $C_{d(k,k')} = (C_{d(k,k')}(1), C_{d(k,k')}(2))$ be the matrix that consists of all unique nonzero vectors in $B_{d(k,k')}$.[19] Then clearly $C_{d(k,k')}$ has full column rank and $C^T_{d(k,k')} C_{d(k,k')}$ has full rank. Thus, $(C^T_{d(k,k')} C_{d(k,k')})^{-1}$ exists. Let $D_{d(k,k')} = (b_{d(k,k')}(1)^T, b_{d(k,k')}(2)^T)^T$. Since by the definition of $b_{d(k,k')}(t)$, $b_{d(k,k')}(t) \cdot b_{d(k,k')}(t')^T = 0$ for $t \neq t'$, $D_{d(k,k')}$ has full row rank and $(D_{d(k,k')} D^T_{d(k,k')})^{-1}$ exists. We then decompose $B_{d(k,k')} = C_{d(k,k')} \cdot D_{d(k,k')}$.[20]

Now by similar proof of Lemma L-17 of Heckman and Pinto (2018), we can show that the Moore–Penrose pseudo inverse of $B_{d(k,k')}$ is

$$B^{+}_{d(k,k')} = D^T_{d(k,k')} (D_{d(k,k')} D^T_{d(k,k')})^{-1} (C^T_{d(k,k')} C_{d(k,k')})^{-1} C^T_{d(k,k')}.$$

For $t \in \{1, 2\}$, we can write $b_{d(k,k')}(t) = e_t D_{d(k,k')}$, where e_t is a row vector in which the tth element is 1 and the other element is 0. Then we have that

$$\begin{aligned}
&b_{d(k,k')}(t)\left[I - B^{+}_{d(k,k')} B_{d(k,k')}\right] = b_{d(k,k')}(t) - b_{d(k,k')}(t) B^{+}_{d(k,k')} B_{d(k,k')} \\
&= b_{d(k,k')}(t) - e_t D_{d(k,k')} D^T_{d(k,k')} (D_{d(k,k')} D^T_{d(k,k')})^{-1} (C^T_{d(k,k')} C_{d(k,k')})^{-1} C^T_{d(k,k')} C_{d(k,k')} \cdot D_{d(k,k')} \\
&= 0.
\end{aligned}$$

[19] Without loss of generality, we assume that both $C_{d(k,k')}(1)$ and $C_{d(k,k')}(2)$ exist.
[20] See Remark A.3 of Heckman and Pinto (2018).

This implies that $b_{d(k,k')}(t) P_{S(k,k')}$ and $b_{d(k,k')}(t) Q_{S(k,k')}(d)$ can be identified by

$$b_{d(k,k')}(t) P_{S(k,k')} = b_{d(k,k')}(t) B^{+}_{d(k,k')} P_{Z(k,k')}(d)$$
$$\text{and } b_{d(k,k')}(t) Q_{S(k,k')}(d) = b_{d(k,k')}(t) B^{+}_{d(k,k')} Q_{Z(k,k')}(d).$$

Thus, (C.2) and (C.3) show that

$$\mathbb{P}\left(\mathcal{M}_{(k,k')} S \in \Sigma_{d(k,k')}(t)\right) = b_{d(k,k')}(t) B^{+}_{d(k,k')} P_{Z(k,k')}(d)$$
$$\text{and } E\left[\kappa\left(Y_d(z_k, z_{k'})\right) | \mathcal{M}_{(k,k')} S \in \Sigma_{d(k,k')}(t)\right] = \frac{b_{d(k,k')}(t) B^{+}_{d(k,k')} Q_{Z(k,k')}(d)}{b_{d(k,k')}(t) B^{+}_{d(k,k')} P_{Z(k,k')}(d)}$$

are identified. Define

$$Z_{Pi} = (1\{Z_i = z_1\}, \ldots, 1\{Z_i = z_K\}),$$
$$P_{DZi}(d) = (1\{D_i = d, Z_i = z_1\}, \ldots, 1\{D_i = d, Z_i = z_K\})^T \text{ for all } d,$$
$$Q_{YDZi}(d) = (\kappa(Y_i) 1\{D_i = d, Z_i = z_1\}, \ldots, \kappa(Y_i) 1\{D_i = d, Z_i = z_K\})^T \text{ for all } d,$$

and

$$W_i = \left(Z_{Pi}, P_{DZi}(d_1)^T, \ldots, P_{DZi}(d_J)^T, Q_{YDZi}(d_1)^T, \ldots, Q_{YDZi}(d_J)^T\right)^T.$$

By multivariate central limit theorem, $\sqrt{n}(\widehat{W} - W) \xrightarrow{d} N(0, \Sigma_W)$, where

$$\Sigma_W = E[(W_i - W)(W_i - W)^T], \tag{C.4}$$

and therefore $\widehat{W} \xrightarrow{p} W$. Also, for every $\varepsilon > 0$, $\mathbb{P}(\sqrt{n}\|\mathbb{1}(\widehat{\mathscr{Z}_0}) - \mathbb{1}(\mathscr{Z}_{\bar{M}})\|_2 > \varepsilon) \leq \mathbb{P}(\widehat{\mathscr{Z}_0} \neq \mathscr{Z}_{\bar{M}}) \to 0$ by assumption. Then, by Lemma 1.10.2(iii) and Example 1.4.7 (Slutsky's lemma) of van der Vaart and Wellner (1996),

$$\sqrt{n}\left\{\left(\widehat{W}^T, \mathbb{1}(\widehat{\mathscr{Z}_0})^T\right)^T - \left(W^T, \mathbb{1}(\mathscr{Z}_{\bar{M}})^T\right)^T\right\} \xrightarrow{d} \left(N(0, \Sigma_W)^T, 0^T\right)^T.$$

■

Proof of Lemma A.3. If $(z_k, z_{k'}) \in \mathscr{Z}_{\bar{M}}$ and $\Sigma_{d(k,k')}(t) = \Sigma_{d'(k,k')}(t')$, then $Y_{dz_k} = Y_d(z_k, z_{k'})$ a.s. and $Y_{d'z_{k'}} = Y_{d'}(z_k, z_{k'})$ a.s. By (A.7), it follows that

$$\beta_{(k,k')}(d, d', t, t') = \left\{\frac{b_{d(k,k')}(t) B^{+}_{d(k,k')} Q_{Z(k,k')}(d)}{b_{d(k,k')}(t) B^{+}_{d(k,k')} P_{Z(k,k')}(d)} - \frac{b_{d'(k,k')}(t') B^{+}_{d'(k,k')} Q_{Z(k,k')}(d')}{b_{d'(k,k')}(t') B^{+}_{d'(k,k')} P_{Z(k,k')}(d')}\right\}. \tag{C.5}$$

If $(z_k, z_{k'}) \notin \mathscr{Z}_{\bar{M}}$ or $\Sigma_{d(k,k')}(t) \neq \Sigma_{d'(k,k')}(t')$, clearly the lemma holds. ■

Proof of Theorem A.4. The proof is similar to that of Theorem A.2. ■

C.2 Definition and Estimation of $\mathscr{Z}_0$

C.2.1 Definition and Estimation of $\mathscr{Z}_1$

Following Sun (2021), we provide the definitions of $\mathscr{Z}_1$ and its estimator. Suppose the instrument Z is pairwise valid with $\mathscr{Z}_{\bar{M}} = \{(z_{k_1}, z_{k'_1}), \ldots, (z_{k_{\bar{M}}}, z_{k'_{\bar{M}}})\}$. Fix $(z, z') \in \mathscr{Z}_{\bar{M}}$. For every $d \in \mathcal{D}$, if $1\{D_{z'} = d\} \leq 1\{D_z = d\}$ a.s., we have that

$$\begin{aligned}\mathbb{P}(Y \in B, D = d | Z = z') &= E\left[1\{Y_d(z, z') \in B\} \times 1\{D_{z'} = d\}\right] \\ &\leq E\left[1\{Y_d(z, z') \in B\} \times 1\{D_z = d\}\right] = \mathbb{P}\left(Y \in B, D = d | Z = z\right) \quad \text{(C.6)}\end{aligned}$$

for all Borel sets B. Denote 2^J J-dimensional different binary vectors by $v_1, \ldots, v_{2^J}$, where

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

Let $\mathcal{L} : \mathcal{D} \to \{1, \ldots, J\}$ map $d \in \mathcal{D}$ to d's index in $\mathcal{D}$ so that if $d = d_j$, we have $\mathcal{L}(d) = j$. For every $q \in \{1, \ldots, 2^J\}$, define $f_q : \{d_1, \ldots, d_J\} \to \{1, -1\}$ by $f_q(d) = (-1)^{v_q(\mathcal{L}(d))}$. For every fixed $(z, z') \in \mathscr{Z}_{\bar{M}}$, there is $q \in \{1, \ldots, 2^J\}$ such that

$$f_q(d) \cdot \{\mathbb{P}\left(Y \in B, D = d | Z = z'\right) - \mathbb{P}\left(Y \in B, D = d | Z = z\right)\} \leq 0$$

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

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

Furthermore, define the following function spaces

$$\mathrm{G} = \left\{\left(1_{\mathbb{R} \times \mathbb{R} \times \{z_j\}}, 1_{\mathbb{R} \times \mathbb{R} \times \{z_k\}}\right) : j, k \in \{1, \ldots, K\}, j < k\right\}, \mathrm{H} = \cup_{q=1}^{2^J} \mathrm{H}_q, \text{ and } \bar{\mathrm{H}} = \cup_{q=1}^{2^J} \bar{\mathrm{H}}_q. \quad \text{(C.7)}$$

Let P and $\widehat{P}$ be defined as in Section 4. Let ϕ, σ^2, $\widehat{\phi}$, and $\widehat{\sigma}^2$ be defined in a way similar to that in Section 4 but for all $(h,g) \in \bar{\mathrm{H}} \times \mathrm{G}$. Also, we let $\Lambda(P) = \prod_{k=1}^{K} P(1_{\mathbb{R}\times\mathbb{R}\times\{z_k\}})$ and $T_n = n \cdot \prod_{k=1}^{K} \widehat{P}(1_{\mathbb{R}\times\mathbb{R}\times\{z_k\}})$. By similar proof of Lemma 3.1 in Sun (2021), σ^2 and $\widehat{\sigma}^2$ are uniformly bounded in $(h,g) \in \bar{\mathrm{H}} \times \mathrm{G}$.

The following lemma reformulates the testable restrictions in terms of ϕ.

Lemma C.1 *Suppose that the instrument Z is pairwise valid for the treatment D with the largest validity pair set $\mathscr{Z}_{\bar{M}} = \{(z_{k_1}, z_{k'_1}), \ldots, (z_{k_{\bar{M}}}, z_{k'_{\bar{M}}})\}$. For every $m \in \{1, \ldots, \bar{M}\}$, we have that $\min_{q\in\{1,\ldots,2^J\}} \sup_{h\in \mathrm{H}_q} \phi(h,g) = 0$ with $g = (1_{\mathbb{R}\times\mathbb{R}\times\{z_{k_m}\}}, 1_{\mathbb{R}\times\mathbb{R}\times\{z_{k'_m}\}})$.*

Proof of Lemma C.1. Since we can find $a \in \mathbb{R}$ and $d \in \mathcal{D}$ such that $P\left(1_{\{a\}\times\{d\}\times\mathbb{R}}\right) = 0$, then we have $\sup_{h\in\mathrm{H}_q} \phi(h,g) \geq 0$ for every q and every $g \in \mathrm{G}$. So for every $g \in \mathrm{G}$, $\min_{q\in\{1,\ldots,2^J\}} \sup_{h\in\mathrm{H}_q} \phi(h,g) \geq 0$. Let $h_{Bd} = 1_{B\times\{d\}\times\mathbb{R}}$ for every closed interval B and every $d \in \mathcal{D}$. Fix $m \in \{1, \ldots, \bar{M}\}$. Under assumption, for every $d \in \mathcal{D}$, we have

$$\phi(h_{Bd}, g) = \frac{P(h_{Bd} \cdot g_2)}{P(g_2)} - \frac{P(h_{Bd} \cdot g_1)}{P(g_1)} \leq 0 \text{ for every closed interval } B,$$
$$\text{or } \phi(-h_{Bd}, g) = \frac{-P(h_{Bd} \cdot g_2)}{P(g_2)} - \frac{-P(h_{Bd} \cdot g_1)}{P(g_1)} \leq 0 \text{ for every closed interval } B,$$

where $g_1 = 1_{\mathbb{R}\times\mathbb{R}\times\{z_{k_m}\}}$, $g_2 = 1_{\mathbb{R}\times\mathbb{R}\times\{z_{k'_m}\}}$, and $g = (g_1, g_2)$. This implies that there is H_q such that $\sup_{h\in\mathrm{H}_q} \phi(h,g) \leq 0$. Thus, it follows that $\min_{q\in\{1,\ldots,2^J\}} \sup_{h\in\mathrm{H}_q} \phi(h,g) = 0$. ■

By Lemma C.1, we define

$$\mathrm{G}_1 = \left\{ g \in \mathrm{G} : \min_{q\in\{1,\ldots,2^J\}} \sup_{h\in\mathrm{H}_q} \phi(h,g) = 0 \right\} \text{ and}$$
$$\widehat{\mathrm{G}_1} = \left\{ g \in \mathrm{G} : \sqrt{T_n} \left| \min_{q\in\{1,\ldots,2^J\}} \sup_{h\in\mathrm{H}_q} \frac{\widehat{\phi}(h,g)}{\xi_0 \vee \widehat{\sigma}(h,g)} \right| \leq \tau_n \right\} \tag{C.8}$$

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 define $\mathscr{Z}_1$ as the collection of all (z, z') that are associated with some $g \in \mathrm{G}_1$:

$$\mathscr{Z}_1 = \left\{ (z_k, z_{k'}) \in \mathscr{Z} : g = (1_{\mathbb{R}\times\mathbb{R}\times\{z_k\}}, 1_{\mathbb{R}\times\mathbb{R}\times\{z_{k'}\}}) \in \mathrm{G}_1 \right\}. \tag{C.9}$$

We use $\widehat{\mathrm{G}_1}$ to construct the estimator of $\mathscr{Z}_1$, denoted by $\widehat{\mathscr{Z}_1}$, which is defined as the set of all (z, z') that are associated with some $g \in \widehat{\mathrm{G}_1}$ in the same way $\mathscr{Z}_1$ is defined based on G_1:

$$\widehat{\mathscr{Z}_1} = \left\{ (z_k, z_{k'}) \in \mathscr{Z} : g = (1_{\mathbb{R}\times\mathbb{R}\times\{z_k\}}, 1_{\mathbb{R}\times\mathbb{R}\times\{z_{k'}\}}) \in \widehat{\mathrm{G}_1} \right\}. \tag{C.10}$$

To derive the desired consistency result, we state and prove an additional auxiliary lemma.

Lemma C.2 *Under Assumption A.6, $\widehat{\phi} \to \phi$, $T_n/n \to \Lambda(P)$, and $\widehat{\sigma} \to \sigma$ almost uniformly. In addition, $\sqrt{T_n}(\widehat{\phi} - \phi) \rightsquigarrow \mathbb{G}$ for some random element $\mathbb{G}$, and for all $(h, g) \in \bar{\mathrm{H}} \times \mathrm{G}$ with $g = (g_1, g_2)$, the variance $Var\left(\mathbb{G}\left(h, g\right)\right) = \sigma^2(h, g)$.*

Proof of Lemma C.2. Note that the spaces $\bar{\mathrm{H}}$ and G defined in (C.7) are similar to the spaces $\bar{\mathcal{H}}$ and $\mathcal{G}_P$ defined in (B.13). The lemma can be proved following a strategy similar to that of the proof of Lemma B.3. ■

Proposition C.1 *Suppose the instrument Z is pairwise valid for the treatment D as defined in Definition A.2. Under Assumption A.6, $\mathbb{P}(\widehat{\mathrm{G}_1} = \mathrm{G}_1) \to 1$, and thus $\mathbb{P}(\widehat{\mathscr{Z}_1} = \mathscr{Z}_1) \to 1$.*

Proof of Proposition C.1. First, suppose $\mathrm{G}_1 \neq \varnothing$. Then we have that

$$\min_{q \in \{1, \ldots, 2^J\}} \sup_{h \in \mathrm{H}_q} \left\{ \phi\left(h, g\right) / \left(\xi_0 \vee \widehat{\sigma}\left(h, g\right)\right) \right\} = 0$$

for all $g \in \mathrm{G}_1$. Under the constructions, we have that

$$\begin{aligned}
&\lim_{n \to \infty} \mathbb{P}\left(\mathrm{G}_1 \setminus \widehat{\mathrm{G}_1} \neq \varnothing \right) \\
&\leq \lim_{n \to \infty} \mathbb{P}\left(\max_{g \in \mathrm{G}_1} \sqrt{T_n} \left| \begin{array}{l} \min_{q \in \{1, \ldots, 2^J\}} \sup_{h \in \mathrm{H}_q} \frac{\widehat{\phi}(h,g)}{\xi_0 \vee \widehat{\sigma}(h,g)} \\ - \min_{q \in \{1, \ldots, 2^J\}} \sup_{h \in \mathrm{H}_q} \frac{\phi(h,g)}{\xi_0 \vee \widehat{\sigma}(h,g)} \end{array} \right| > \tau_n \right) \\
&= \lim_{n \to \infty} \mathbb{P}\left(\max_{g \in \mathrm{G}_1} \sqrt{T_n} \left| \begin{array}{l} - \max_{q \in \{1, \ldots, 2^J\}} \left(- \sup_{h \in \mathrm{H}_q} \frac{\widehat{\phi}(h,g)}{\xi_0 \vee \widehat{\sigma}(h,g)} \right) \\ + \max_{q \in \{1, \ldots, 2^J\}} \left(- \sup_{h \in \mathrm{H}_q} \frac{\phi(h,g)}{\xi_0 \vee \widehat{\sigma}(h,g)} \right) \end{array} \right| > \tau_n \right) \\
&\leq \lim_{n \to \infty} \mathbb{P}\left(\max_{g \in \mathrm{G}_1} \sup_{h \in \mathrm{H}} \sqrt{T_n} \left| \frac{\widehat{\phi}\left(h, g\right) - \phi\left(h, g\right)}{\xi_0 \vee \widehat{\sigma}\left(h, g\right)} \right| > \tau_n \right).
\end{aligned}$$

By Lemma C.2, $\sqrt{T_n}(\widehat{\phi} - \phi) \rightsquigarrow \mathbb{G}$ and $\widehat{\sigma} \to \sigma$ almost uniformly, which implies that $\widehat{\sigma} \rightsquigarrow \sigma$ by Lemmas 1.9.3(ii) and 1.10.2(iii) of van der Vaart and Wellner (1996). Then by Example 1.4.7 (Slutsky's lemma) and Theorem 1.3.6 (continuous mapping) of van der Vaart and Wellner (1996),

$$\max_{g \in \mathrm{G}_1} \sup_{h \in \mathrm{H}} \sqrt{T_n} \left| \frac{\widehat{\phi}\left(h, g\right) - \phi\left(h, g\right)}{\xi_0 \vee \widehat{\sigma}\left(h, g\right)} \right| \rightsquigarrow \max_{g \in \mathrm{G}_1} \sup_{h \in \mathrm{H}} \left| \frac{\mathbb{G}\left(h, g\right)}{\xi_0 \vee \sigma\left(h, g\right)} \right|.$$

Since $\tau_n \to \infty$, we have that $\lim_{n\to\infty} \mathbb{P}(\mathrm{G}_1 \setminus \widehat{\mathrm{G}_1} \neq \varnothing) = 0$.

If $\mathrm{G}_1 = \mathrm{G}$, then clearly $\lim_{n\to\infty} \mathbb{P}(\widehat{\mathrm{G}_1} \setminus \mathrm{G}_1 \neq \varnothing) = 0$. Suppose now $\mathrm{G}_1 \neq \mathrm{G}$. Since G is a finite set and $\widehat{\sigma}$ is uniformly bounded, then there is a $\delta > 0$ such that

$$\min_{g\in \mathrm{G}\setminus \mathrm{G}_1} \left| \min_{q\in\{1,\ldots,2^J\}} \sup_{h\in \mathrm{H}_q} \frac{\phi(h,g)}{\xi_0 \vee \widehat{\sigma}(h,g)} \right| > \delta.$$

By Lemma C.2, $\widehat{\phi} \to \phi$ almost uniformly. Thus, for every $\varepsilon > 0$, there is a measurable set A with $\mathbb{P}(A) \geq 1 - \varepsilon$ such that for sufficiently large n,

$$\max_{g\in \mathrm{G}} \left| \left| \min_{q\in\{1,\ldots,2^J\}} \sup_{h\in \mathrm{H}_q} \frac{\widehat{\phi}(h,g)}{\xi_0 \vee \widehat{\sigma}(h,g)} \right| - \left| \min_{q\in\{1,\ldots,2^J\}} \sup_{h\in \mathrm{H}_q} \frac{\phi(h,g)}{\xi_0 \vee \widehat{\sigma}(h,g)} \right| \right| \leq \frac{\delta}{2} \tag{C.11}$$

uniformly on A. We now have that

$$\begin{aligned}
&\lim_{n\to\infty} \mathbb{P}\left(\widehat{\mathrm{G}_1} \setminus \mathrm{G}_1 \neq \varnothing\right) \\
&\leq \lim_{n\to\infty} \mathbb{P}\left(\begin{array}{c} \left\{ \max_{g\in \widehat{\mathrm{G}_1}\setminus \mathrm{G}_1} \left| \min_{q\in\{1,\ldots,2^J\}} \sup_{h\in \mathrm{H}_q} \frac{\phi(h,g)}{\xi_0 \vee \widehat{\sigma}(h,g)} \right| > \delta \right\} \\ \cap \left\{ \max_{g\in \widehat{\mathrm{G}_1}\setminus \mathrm{G}_1} \sqrt{T_n} \left| \min_{q\in\{1,\ldots,2^J\}} \sup_{h\in \mathrm{H}_q} \frac{\widehat{\phi}(h,g)}{\xi_0 \vee \widehat{\sigma}(h,g)} \right| \leq \tau_n \right\} \cap A \end{array} \right) + \mathbb{P}(A^c) \\
&\leq \lim_{n\to\infty} \mathbb{P}\left(\sqrt{\frac{T_n}{n}} \frac{\delta}{2} < \max_{g\in \widehat{\mathrm{G}_1}\setminus \mathrm{G}_1} \sqrt{\frac{T_n}{n}} \left| \min_{q\in\{1,\ldots,2^J\}} \sup_{h\in \mathrm{H}_q} \frac{\widehat{\phi}(h,g)}{\xi_0 \vee \widehat{\sigma}(h,g)} \right| \leq \frac{\tau_n}{\sqrt{n}} \right) + \varepsilon = \varepsilon,
\end{aligned}$$

because $\tau_n/\sqrt{n} \to 0$ as $n \to \infty$. Here, ε can be arbitrarily small. Thus we have that $\mathbb{P}(\widehat{\mathrm{G}_1} = \mathrm{G}_1) \to 1$, because $\mathbb{P}(\mathrm{G}_1 \setminus \widehat{\mathrm{G}_1} \neq \varnothing) \to 0$ and $\mathbb{P}(\widehat{\mathrm{G}_1} \setminus \mathrm{G}_1 \neq \varnothing) \to 0$.

Second, suppose $\mathrm{G}_1 = \varnothing$. This implies that

$$\min_{g\in \mathrm{G}} \left| \min_{q\in\{1,\ldots,2^J\}} \sup_{h\in \mathrm{H}_q} \frac{\phi(h,g)}{\xi_0 \vee \widehat{\sigma}(h,g)} \right| > \delta$$

for some $\delta > 0$. Thus, with (C.11) we now have that

$$\begin{aligned}
&\lim_{n\to\infty} \mathbb{P}\left(\widehat{\mathrm{G}_1} \neq \varnothing\right) \\
&\leq \lim_{n\to\infty} \mathbb{P}\left(\begin{array}{c} \left\{ \max_{g\in \widehat{\mathrm{G}_1}} \left| \min_{q\in\{1,\ldots,2^J\}} \sup_{h\in \mathrm{H}_q} \frac{\phi(h,g)}{\xi_0 \vee \widehat{\sigma}(h,g)} \right| > \delta \right\} \\ \cap \left\{ \max_{g\in \widehat{\mathrm{G}_1}} \sqrt{T_n} \left| \min_{q\in\{1,\ldots,2^J\}} \sup_{h\in \mathrm{H}_q} \frac{\widehat{\phi}(h,g)}{\xi_0 \vee \widehat{\sigma}(h,g)} \right| \leq \tau_n \right\} \cap A \end{array} \right) + \mathbb{P}(A^c) \\
&\leq \lim_{n\to\infty} \mathbb{P}\left(\sqrt{\frac{T_n}{n}} \frac{\delta}{2} < \max_{g\in \widehat{\mathrm{G}_1}} \sqrt{\frac{T_n}{n}} \left| \min_{q\in\{1,\ldots,2^J\}} \sup_{h\in \mathrm{H}_q} \frac{\widehat{\phi}(h,g)}{\xi_0 \vee \widehat{\sigma}(h,g)} \right| \leq \frac{\tau_n}{\sqrt{n}} \right) + \varepsilon = \varepsilon,
\end{aligned}$$

because $\tau_n/\sqrt{n} \to 0$ as $n \to \infty$. Here, ε can be arbitrarily small. Thus, $\mathbb{P}(\widehat{\mathrm{G}_1} = \mathrm{G}_1) = 1 - \mathbb{P}(\widehat{\mathrm{G}_1} \neq \varnothing) \to 1$. ■

Proposition C.1 is also related to the contact set estimation in Sun (2021). Since G is a finite set, we can obtain the stronger result in Proposition C.1, that is, $\mathbb{P}(\widehat{\mathrm{G}_1} = \mathrm{G}_1) \to 1$.

C.2.2 Definition and Estimation of $\mathscr{Z}_2$

The definition of $\mathscr{Z}_2$ is the same as that in Appendix B.4.2 because the necessary conditions provided by Kédagni and Mourifié (2020) are for the exclusion and statistical independence conditions only. Therefore, the estimator of $\mathscr{Z}_2$ can be constructed as in Section B.4.2.