A Test for Instrument Validity*

Toru Kitagawa†

CeMMAP and *Department of Economics, UCL*

This draft: June, 2015

Abstract

This paper develops a specification test for instrument validity in the heterogeneous treatment effect model with a binary treatment and a discrete instrument. The strongest testable implication for instrument validity is given by the condition for non-negativity of point-identifiable complier's outcome densities. Our specification test infers this testable implication using a variance-weighted Kolmogorov-Smirnov test statistic. The test can be applied to both discrete and continuous outcome cases, and an extension of the test to settings with conditioning covariates is provided.

Keywords: Treatment Effects, Instrumental Variable, Specification Test, Bootstrap.
JEL Classification: C12, C15, C21.

*This paper is a revised version of a chapter of my Ph.D thesis submitted to Brown University in 2009. This paper replaces the previous versions titled as "A Bootstrap Test for Instrument Validity in the Heterogeneous Treatment Effect Models."

†Email: t.kitagawa@ucl.ac.uk. I would like to thank Frank Kleibergen for his guidance and continuous encouragement. I would also like to thank Josh Angrist, Tim Armstrong, Clément de Chaisemartin, Le-Yu Chen, Mario Fiorini, James Heckman, Stefan Hoderlein, Yu-Chin Hsu, Martin Huber, Hide Ichimura, Guido Imbens, Giovanni Mellace, and Katrine Stevens, and the seminar participants at Academia Sinica, Collegio Carlo Alberto, GRIPS, Pennsylvania State University, Simon Fraser University, Stockholm University, and UCL, for helpful comments and beneficial discussions. I also thank a co-editor and three anonymous referees for valuable suggestions that significantly improved the paper. Financial support from the ESRC through the ESRC Center for Microdata Methods and Practice (CEMMAP) (grant number RES-589-28-0001) and the Merit Dissertation Fellowship from the Graduate School of Economics in Brown University are gratefully acknowledged.

1 Introduction

Consider a heterogeneous causal effect model of Angrist and Imbens (1994) with a binary treatment and a binary instrument. We denote an observed outcome by $Y \in \mathcal{Y} \subset \mathbb{R}$, an observed treatment status by $D \in \{1,0\}$; $D = 1$ when one receives the treatment while $D = 0$ when one does not, and a binary non-degenerate instrument by $Z \in \{1,0\}$. Let $\{Y_{dz} \in \mathcal{Y} : d \in \{1,0\}, z \in \{1,0\}\}$ be the potential outcomes that would have been observed if the treatment status were set at $D = d$ and the assigned instrument were set at $Z = z$. Furthermore, $\{D_z : z \in \{1,0\}\}$ are the potential treatment responses that would have been observed if $Z = 1$ and $Z = 0$, respectively. The seminal works of Imbens and Angrist (1994) and Angrist, Imbens and Rubin (1996) showed that, given $\Pr(D = 1|Z = 1) > \Pr(D = 1|Z = 0)$, the instrument variable Z that satisfies the three conditions involving the potential variables is able to identify the average treatment effects for those whose selection to treatment is affected by the instrument (local average treatment effect, LATE hereafter). The three key conditions, of which the joint validity is hereafter referred to as *IV-validity,* are[1]

Assumption: IV-validity for binary Z

(i) *Instrument Exclusion*: $Y_{d1} = Y_{d0}$ for $d = 1, 0$, with probability one.

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

(iii) *Instrument Monotonicity (No-defier):* The potential treatment response indicators satisfy $D_1 \geq D_0$ with probability one.

Despite the fact that the credibility of LATE analysis relies on the validity of the employed instrument, no test procedure has been proposed to empirically diagnose IV-validity. As a result, causal inference studies have assumed IV-validity based solely on some background knowledge or out-of-sample evidence, and, accordingly, its credibility often remains controversial in many empirical contexts.

The main contribution of this paper is to develop a specification test for IV-validity in the LATE model. Our specification test builds on the testable implication obtained by Balke

[1]Note that the null hypothesis of IV-validity tested in this paper does not include the *instrument relevance* assumption, $\Pr(D = 1|Z = 1) > \Pr(D = 1|Z = 0)$. The instrument relevance assumption can be assessed by inferring the coefficient in the first-stage regression of D onto Z.

and Pearl (1997) and Heckman and Vytlacil (2005, Proposition A.5). Let P and Q be the conditional probability distributions of $(Y, D) \in \mathcal{Y} \times \{1, 0\}$ given $Z = 1$ and $Z = 0$, i.e.,

$$\begin{aligned} P(B, d) &= \Pr(Y \in B, D = d | Z = 1), \\ Q(B, d) &= \Pr(Y \in B, D = d | Z = 0), \end{aligned}$$

for Borel set B in $\mathcal{Y}$ and $d = 1, 0$. Since P and Q are conditional distribution of observable variables, they are identified by the sampling process. Imbens and Rubin (1997) showed that, under IV-validity,

$$\begin{aligned} P(B, 1) - Q(B, 1) &= \Pr(Y_1 \in B, D_1 > D_0) \text{ and} \\ Q(B, 0) - P(B, 0) &= \Pr(Y_0 \in B, D_1 > D_0) \end{aligned}$$

hold for every B in $\mathcal{Y}$. Since the quantities in the right-hand sides are nonnegative by the definition of probabilities, we obtain the testable implication of Balke and Pearl (1997) and Heckman and Vytlacil (2005);

$$\begin{aligned} P(B, 1) - Q(B, 1) &\geq 0, \\ Q(B, 0) - P(B, 0) &\geq 0, \end{aligned} \tag{1.1}$$

for every Borel set B in $\mathcal{Y}$.[2] Figures 1 and 2 provide visual illustration of these testable implications for a continuous Y case. The solid lines, $p(y, d)$ and $q(y, d)$, plot the probability density of $P(\cdot, d)$ and $Q(\cdot, d)$ over Y-axis at fixed $d \in \{1, 0\}$. It is important to keep in mind that, in the presence of noncompliance, integrations of $p(y, d)$ and $q(y, d)$ over $y \in \mathcal{Y}$ are smaller than one, as they are equal to $\Pr(D = d | z = 1) < 1$ and $\Pr(D = d | z = 0) < 1$, respectively. If the instrument is valid, $p(y, 1)$ must nest $q(y, 1)$ for treatment outcome, and $q(y, 0)$ must nest $p(y, 0)$ for control outcome, as plotted in Figure 1.

In contrast, if we observe the densities as plotted in Figure 2, we can refute at least one of the IV-validity assumptions since some of the inequalities (1.1) are violated at some subsets in the support of Y, e.g., those labeled as V_1 and V_2 in Figure 2.

To see how densities of $P(\cdot, d)$ and $Q(\cdot, d)$ look like in real data, Figure 3 plots kernel density estimates of $p(y, d)$ and $q(y, d)$ for the Vietnam era draft lottery data used in Angrist and Krueger (1992, 1995) and Abadie (2002), where $Y = \log((\text{one's post-war annual}$

[2] As is clear from the derivation, the testable implication can be equivalently interpreted as the nonnegativity conditions for the complier's potential outcome distributions, $\Pr(Y_d \in B | D_1 > D_0) \geq 0$, which are identifiable under IV-validity. Imbens and Rubin (1997) noted that, depending on data, the estimates of the complier's outcome densities can be negative over some region in the outcome support.

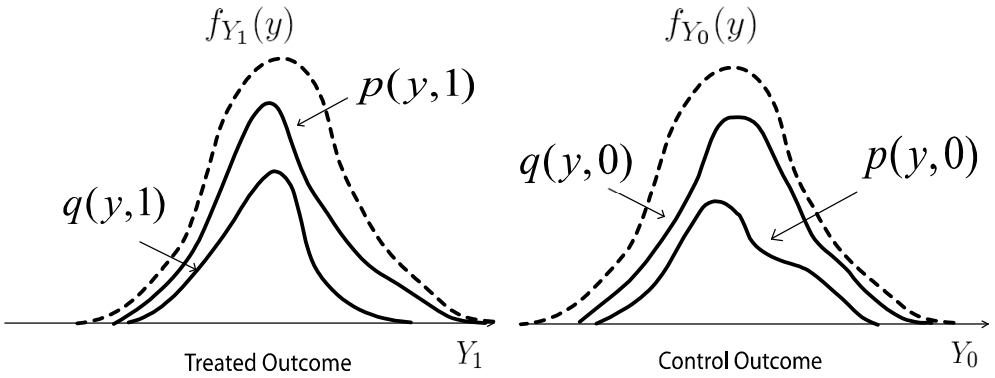

Figure 1: *If the identifiable densities $p(y, D = d)$ and $q(y, D = d)$ are nested as in this figure, IV-validity cannot be refuted. The dotted lines show the marginal probability densities of the potential outcomes, i.e., $f_{Y_d}(y)$ is the marginal probability density of $Y_d \equiv Y_{d1} = Y_{d0}$, which is not identifiable. Under the instrument exclusion and random assignment, both $p(y, d)$ and $q(y, d)$ must lie below the potential outcome densities $f_{Y_d}(\cdot)$.*

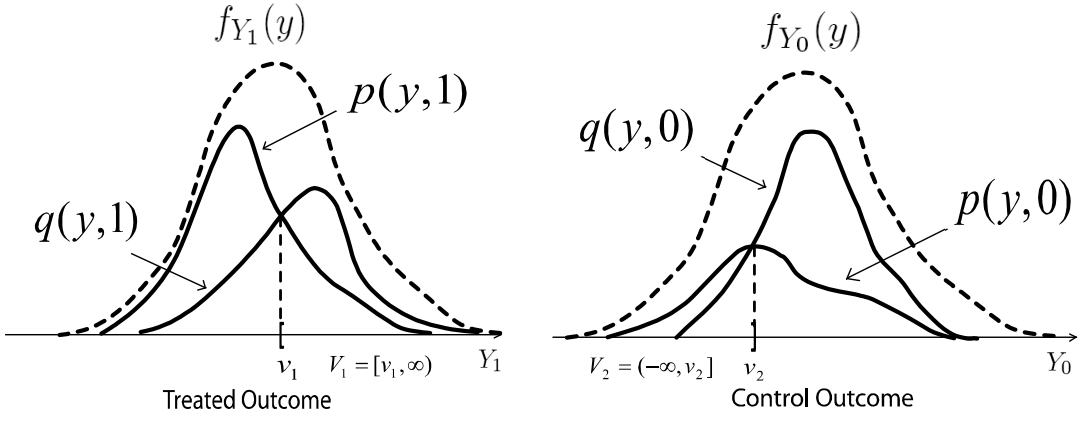

Figure 2: *If $p(y, d)$ intersects with $q(y, d)$ for at least one of $d = 1, 0$, we can refute IV-validity.*

earnings)+1), the veteran status is D, and the draft eligibility determined by a low lottery number is Z. See Section 4.1 for the detail description of the data. The estimated densities overall exhibit the nesting relationships similar to those illustrated in Figure 1;[3] therefore, no strong evidence against IV-validity appears to be available. Contrasting density plots are shown in Figure 4, where the data are from Card (1993), Y is the logarithm of one's weekly earning, D indicates whether one graduated from a four-year college, Z indicates whether a four-year college is located in the area of one's residence. No conditioning covariates are controlled for when drawing the densities. Here, we observe that the density estimates intersect, especially for the control outcome. This is an in-sample visual evidence against IV-validity. These eyeball-based assessments are indeed intuitive and useful, but they fail (i) to take into account sampling uncertainty and (ii) to quantify the strength of evidence for or against IV-validity without relying on a specific choice of smoothing parameters. A hypothesis test procedure proposed in this paper solves these important practical issues.

The above derivation of (1.1) shows only that inequalities (1.1) are *necessary* implications of IV-validity, so it is natural to ask (i) whether the testable implications of (1.1) can be further strengthened and (ii) whether there exist some P and Q for which (1.1) becomes a necessary and sufficient condition for IV-validity. The next proposition shows that the answers to these questions are negative (see Kitagawa (2015, Appendix A) for a proof).

Proposition 1.1 *Assume that $P(\cdot, d)$ and $Q(\cdot, d)$ have a common dominating measure μ on $\mathcal{Y}$ for each $d = 1, 0$. (i) If distributions of observables, P and Q, satisfy inequalities (1.1), then there exists a joint probability law of $(Y_{11}, Y_{10}, Y_{01}, Y_{00}, D_1, D_0, Z)$ that satisfies IV-validity and induces the P and Q.*

(ii) For any P and Q satisfying inequalities (1.1), we can construct a joint probability law of $(Y_{11}, Y_{10}, Y_{01}, Y_{00}, D_1, D_0, Z)$ that violates IV-validity.

To my knowledge, Proposition 1.1 is new in the literature, and it shows the following important results. First, Proposition 1.1 (i) shows an optimality of the testable implication (1.1), in the sense that any other feature of the data distribution cannot contribute to

[3] The probability subdensities have probability masses at $Y = 0$, as the data include individuals with zero earnings. The sample estimates of these probability masses satisfy (1.1). Qualitatively similar estimates of the subdensities are obtained if we define the outcome as $Y = \log((\text{weekly wages})+1)$. Our test can be applied without any change even when the distribution of outcome has probability masses.

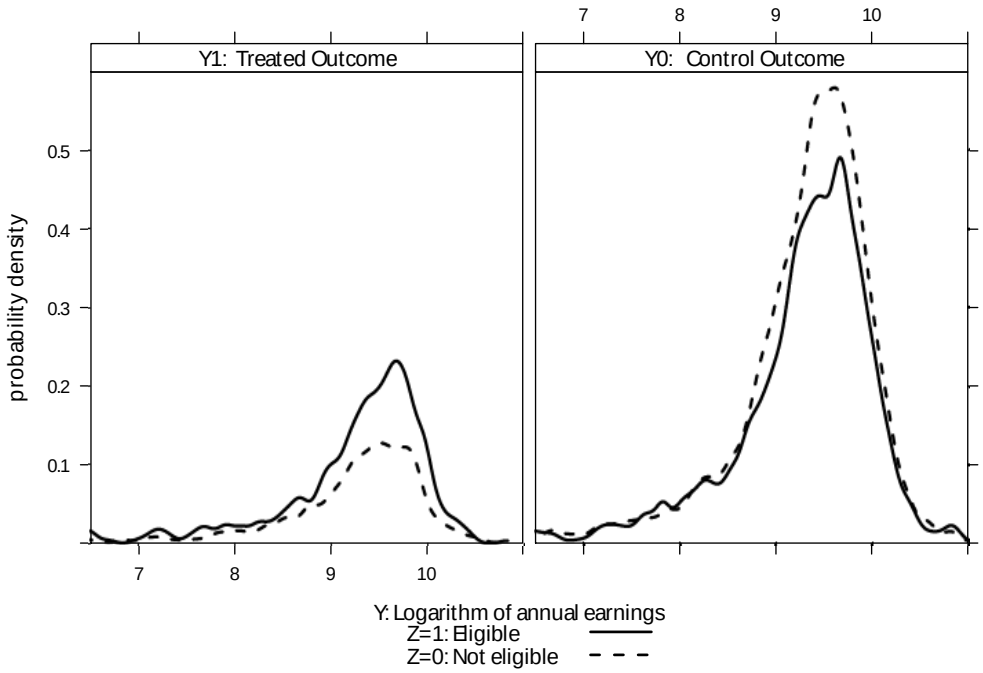

Figure 3: **Kernel Density Estimates, Draft Lottery Data.** The *Gaussian kernel with bandwidth 0.07 is used.*

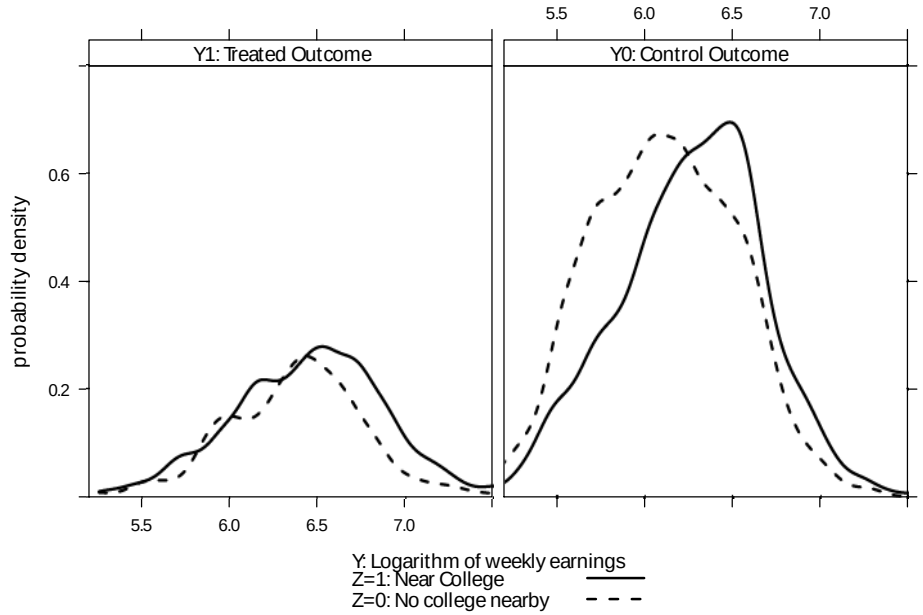

Figure 4: **Kernel Density Estimates, Proximity to College Data**. *The Gaussian kernel with bandwidth 0.08 is used.*

screening out invalid instruments further than the testable implication of (1.1). Second, Proposition 1.1 (ii) highlights limitation on learnability of instrument validity in the sense that accepting the null hypothesis of (1.1) never allows us to *confirm* IV-validity no matter how large the sample size is. In this precise sense, the IV-validity is a *refutable* but *non-verifiable* hypothesis. Such limitation on confirmability of instrument validity is known in other contexts, such as the classical over-identification test in the linear instrumental variable method with homogeneous effect[4] and the test of instrument monotonicity in the multi-valued treatment case proposed in Angrist and Imbens (1995).[5] See Breusch (1986) for a general discussion on hypothesis testing of refutable but non-verifiable assumptions.

Our test uses a variance-weighted Kolmogorov-Smirnov test statistic (KS-statistic, hereafter) to measure the magnitude of violations of inequalities (1.1) in the data. We provide a resampling algorithm to obtain critical values and demonstrate that the test procedure attains asymptotically correct size uniformly over a large class of data generating processes, and consistently rejects all the data generating processes violating (1.1). A similar variance weighted KS-statistic has been considered in the literature of conditional moment inequalities, as in Andrews and Shi (2013), Armstrong (2014), Armstrong and Chan (2013), and Chetverikov (2012). As shown in Romano (1988), bootstrap is widely applicable and easy to implement to obtain the critical values for general KS-statistic, and it has been instrumental in the context of stochastic dominance testing; see, e.g., Abadie (2002), Barrett and Donald (2003), Donald and Hsu (forthcoming), Horváth, Kokoszka, and Zitikis (2006), and Linton, Maasoumi, and Whang (2005).

Our test concerns the exogeneity of instrument defined in terms of statistical independence, and it can be applied to the context in which objects of interest are distributional features of complier's potential outcome distribution, e.g., the quantile treatment effects for

[4]If the instrument is multi-valued, we can naively perform the classical over-identification test by treating the multi-valued instrument as a collection of binary instruments. However, as discussed in Imbens (2014) and Lee (2014), the over-identification test should not be used if causal effects are considered to be heterogeneous, since heterogeneity of causal effects can lead to misspecified over-identifying restrictions, even when LATE IV-validity is true.

[5]In case of multi-valued treatment status, Angrist and Imbens (1995) propose a specification test to assess instrument monotonicity by inferring the stochastic dominance of the distribution functions of the treatment status conditional on the instrument; see Barua and Lang (2009) and Fiorini, Stevens, Taylor, and Edwards (2013) for applications of the Angrist and Imbens test. In the binary treatment case, however, Angrist and Imbens test cannot be applied.

compliers (Abadie, Angrist, and Imbens (2002)). On the other hand, if solely the mean effect is concerned, identification of LATE can in fact be attained under a slightly weaker set of assumptions, such that the instrument is statistically independent of the selection types while the potential outcomes are only mean independent of Z conditional on each selection type. Huber and Mellace (2013) show that this weaker LATE identifying condition has a testable implication given by a finite number of moment inequalities. Since our test builds on the distributional restrictions implied from statistical independence, it screens out a larger class of data generating processes compared to the test of Huber and Mellace. In addition, the set of detectable alternatives and the p-value of our test are invariant to any monotonic transformation of the outcome variables, whereas this invariance property does not hold for the Huber and Mellace's test. Mourifié and Wan (2014) recently proposes an alternative way to test the same instrument validity condition by transforming the testable implication (1.1) into conditional moment inequality restrictions. For the binary Y case, Machado, Shaikh, and Vytlacil (2013) develops a multiple hypothesis testing procedure that jointly infers IV-validity and the sign of average treatment effect.

The rest of the paper is organized as follows. Section 2 presents implementation of our test when D and Z are binary and shows its asymptotic validity. Section 3 extends the analysis to settings with a multi-valued instrument and with conditioning covariates. Two empirical applications are provided in Section 4. The online supplementary material Kitagawa (2015) provides proofs and the results of Monte Carlo experiments.

2 Test

2.1 Test Statistics and Implementation

Let a sample be given by N observations of $(Y, D, Z) \in \mathcal{Y} \times \{1, 0\}^2$. We divide the sample into two subsamples based on the value of Z, and we consider the sampling process as being conditional on a sequence of instrument values. Let (Y_i^1, D_i^1), $i = 1, \ldots, m$ be observations with $Z = 1$ and (Y_j^0, D_j^0), $j = 1, \ldots, n$ be those with $Z = 0$, and assume that the observations of (Y_i^1, D_i^1) and (Y_j^0, D_j^0) are drawn independently and identically from P and Q, respectively. We assume a deterministic sequence $\hat{\lambda} = m/N \to \lambda$ as $N \to \infty$, where $0 < \lambda < 1$.[6] We

[6] If one wants to perform the test without conditioning on observations of Z, instruments need be resampled as well in the bootstrap algorithm given below.

denote the empirical distributions of P and Q by

$$P_m(B,d) \equiv \frac{1}{m}\sum_{i=1}^{m} I\{Y_i^1 \in B, D_i^1 = d\},$$
$$Q_n(B,d) \equiv \frac{1}{n}\sum_{j=1}^{n} I\{Y_j^0 \in B, D_j^0 = d\}.$$

To test the null hypothesis given by inequalities (1.1), we consider a variance-weighted KS-statistic,

$$T_N = \left(\frac{mn}{N}\right)^{1/2} \max\left\{ \begin{array}{c} \sup_{-\infty\le y\le y'\le\infty}\left\{\frac{Q_n([y,y'],1)-P_m([y,y'],1)}{\xi\vee\sigma_{P_m,Q_n}([y,y'],1)}\right\}, \\ \sup_{-\infty\le y\le y'\le\infty}\left\{\frac{P_m([y,y'],0)-Q_n([y,y'],0)}{\xi\vee\sigma_{P_m,Q_n}([y,y'],0)}\right\} \end{array} \right\}, \tag{2.1}$$

where ξ is a positive constant specified by the user and

$$\sigma^2_{P_m,Q_n}([y,y'],d) = (1-\hat{\lambda})P_m([y,y'],d)(1-P_m([y,y'],d)) + \hat{\lambda}Q_n([y,y'],d)(1-Q_n([y,y'],d)).$$

If the sample counterpart of the first (second) inequality of (1.1) is violated at some interval, then, the first (second) supremum in the max operator becomes positive. For each interval $[y,y']$, $\sigma^2_{P_m,Q_n}([y,y'],d)$ is a consistent estimator of the asymptotic variance of $\left(\frac{mn}{N}\right)^{1/2}(P_m([y,y'],d) - Q_n([y,y'],d))$. Thus, the proposed test statistics quantifies a variance-adjusted maximal violation of the inequalities (1.1) over a class of connected intervals including unbounded ones. The exact suprema can be computed by evaluating the maximand at the finite number of intervals, because, to compute the first (second) supremum in the statistic, it suffices to evaluate the differences of the empirical distribution functions at every interval, the boundaries of which are given by a pair of Y values observed in the subsample of $\{D=1, Z=0\}$ ($\{D=0, Z=1\}$). The suprema are searched over a smaller class of subsets than the class of Borel sets for which the population inequalities (1.1) are examined. Nevertheless, this reduction of the class of sets does not cause any loss of information, in the sense that any data generating processes that violate (1.1) for at least one Borel set can be screened out asymptotically (Theorem 2.1 (ii) below). Note that the proposed test statistic and asymptotic validity of the test are not restricted to a continuous Y case. The same statistic can be used for any ordered discrete Y or a mixture of discrete and continuous Y.[7]

[7] A similar test statistic can be defined also for unordered discrete Y and multi-dimensional Y. In case of unordered discrete Y, the supremum can be defined over every support point of Y, and in case of multi-dimensional Y, the supremum can be defined over a class of rectangles in the support of Y.

The user-specified trimming constant ξ plays a role in ensuring that the inverse weighting terms are sufficiently away from zero. Note that when $\xi \geq 1/2$, the proposed test statistic is identical up to a constant to the non-weighted KS-statistic,

$$T_{N,nw} = \left(\frac{mn}{N}\right)^{1/2} \max \left\{ \begin{array}{c} \sup_{-\infty \leq y \leq y' \leq \infty} \left\{ Q_n([y,y'],1) - P_m([y,y'],1) \right\}, \\ \sup_{-\infty \leq y \leq y' \leq \infty} \left\{ P_m([y,y'],0) - Q_n([y,y'],0) \right\} \end{array} \right\}. \tag{2.2}$$

Hence, variance-weighting is effective only when ξ is smaller than $1/2$. The Monte Carlo studies presented in Kitagawa (2015, Appendix D) show that the test size is insensitive to a choice of ξ even in small sample situations. The finite sample power of the test, on the other hand, can be sensitive to a choice of ξ depending on a specification of alternative. Specifically, when violations of the testable implications occur at the tail parts of P and Q, our Monte Carlo experiments suggest that smaller ξ yields a higher power. In contrast, if violations occur at an interval where P and Q have high probabilities, a larger ξ tends to show a slightly higher power. Although a formal discussion regarding an optimal choice of ξ is out of scope of this paper, our informal recommendation is to specify ξ to a value in the range of 0.05 and 0.1 in order to avoid a big power loss when violations are occurring at the tail parts of P and Q. Alternatively, reporting the test results with several choices of ξ is also recommended in order to showcase the range of p-values over different choices of ξ.

To obtain asymptotically valid critical values for the test, we focus on a data generating processes on the boundary of the one-sided null hypothesis, such that P and Q are identical to some probability measure H. Specifically, we set H at the pooled probability measure (the unconditional distribution of (Y,D)),[8]

$$H = \lambda P + (1-\lambda)Q, \tag{2.3}$$

and aim to estimate the quantiles of the null distribution of the statistic as if the data were generated from $P = Q = H$.[9]

[8] Instead of the pooled probability measure, a different convex combination of P and Q,

$$H = cP + (1-c)Q, \quad c \in [0,1],$$

can be used to generate the bootstrap samples without distorting the size property of the test. The power of the test, on the other hand, can vary depending on the choice of c. We leave investigation about a desirable choice of c for future research.

[9] The finite sample power may be improved if critical values are obtained from the null distribution of the supremum statistic over a pre-estimated set of y where $p(y,d) = q(y,d)$ (contact set). See Lee, Song,

We now summarize a bootstrap algorithm for obtaining critical values for T_N.

Algorithm 2.1:

(i) *Sample* (Y_i^*, D_i^*), $i = 1, \ldots, m$ *randomly with replacement from* $H_N = \hat{\lambda} P_m + \left(1 - \hat{\lambda}\right) Q_n$ *and construct empirical distribution* P_m^*. *Similarly, sample* (Y_j^*, D_j^*), $j = 1, \ldots, n$ *randomly with replacement from* H_N *and construct empirical distribution* Q_n^*.

(ii) *Calculate a bootstrap realization of test statistic*[10]

$$T_N^* = \left(\frac{mn}{N}\right)^{1/2} \max \left\{ \begin{array}{c} \sup_{-\infty \leq y \leq y' \leq \infty} \left\{ \frac{Q_n^*([y,y'],1) - P_m^*([y,y'],1)}{\xi \vee \sigma_{P_m^*, Q_n^*}([y,y'],1)} \right\}, \\ \sup_{-\infty \leq y \leq y' \leq \infty} \left\{ \frac{P_m^*([y,y'],0) - Q_n^*([y,y'],0)}{\xi \vee \sigma_{P_m^*, Q_n^*}([y,y'],0)} \right\} \end{array} \right\},$$

where $\sigma^2_{P_m^*, Q_n^*}([y,y'],d) = (1-\hat{\lambda}) P_m^*([y,y'],d)(1 - P_m^*([y,y'],d)) + \hat{\lambda} Q_n^*([y,y'],d)(1 - Q_n^*([y,y'],d))$.

(iii) *Iterate Step 1 - 3 many times and get the empirical distribution of* T_N^*. *For a chosen nominal level* $\alpha \in (0, 1/2)$, *we obtain a bootstrap critical value* $c_{N,1-\alpha}$ *from its empirical* $(1-\alpha)$*-th quantile .*

(iv) *Reject the null hypothesis (1.1) if* $T_N > c_{N,1-\alpha}$. *The bootstrap p-value is obtained according to the proportion of bootstrap repetitions such that* T_N^* exceeds T_N.

2.2 Asymptotic Uniform Size Control and Consistency

This section formally claims that the test procedure of Algorithm 2.1 is asymptotically valid uniformly over a certain class of data generating processes. Let $\mathcal{P}$ be a set of probability measures defined on the Borel σ-algebra of $\mathcal{Y} \times \{0,1\}$, and the set of data generating processes satisfying (1.1) is denoted by

$$\mathcal{H}_0 = \left\{ (P,Q) \in \mathcal{P}^2 : \text{ inequalities (1.1) hold.} \right\}.$$

and Whang (2011), Linton, Song, and Whang (2010)), Donald and Hsu (forthcoming), and the literatures on generalized moment selection including Andrews and Barwick (2012), Andrews and Shi (2013), Andrews and Soares (2010), among others. Estimation of the contact set relies on a user-specified tuning parameter, and the test size can be affected by its choice.

[10] Since P_m^* and Q_n^* are drawn from the common pooled empirical distribution, recentering of the bootstrap empirical measures with respect to the original P_m and Q_n are not needed.

The uniform validity of our test procedure is based on the following two weak regularity conditions.

Condition-RG:
(a) Probability measures in $\mathcal{P}$ are nondegenerate and have a common dominating measure μ for the $\mathcal{Y}$-coordinate, where μ is the Lebesgue measure, a point mass measure with finite support points, or their mixture. The density functions $p(y,d) \equiv \frac{dP(\cdot,d)}{d\mu}$ are bounded uniformly over $\mathcal{P}$, i.e., there exists $M < \infty$ such that $p(y,d) \leq M$ holds at μ-almost every $y \in \mathcal{Y}$ and $d = 0, 1$ for all $P \in \mathcal{P}$.
(b) $\mathcal{P}$ is uniformly tight, i.e., for arbitrary $\epsilon > 0$, there exists a compact set $K \subset \mathcal{Y} \times \{0,1\}$ such that

$$\sup_{P \in \mathcal{P}} \{P(K^c)\} < \epsilon.$$

The asymptotic validity of the proposed test is stated in the next proposition (see Kitagawa (2015, Appendix B) for a proof).

Theorem 2.1 *Let $\alpha \in (0, 1/2)$. (i) Suppose Condition-RG. The test procedure of Algorithm 2.1 has asymptotically uniformly correct size for null hypothesis $\mathcal{H}_0$,*

$$\limsup_{N \to \infty} \sup_{(P,Q) \in \mathcal{H}_0} \Pr\left(T_N > c_{N,1-\alpha}\right) \leq \alpha.$$

(ii) For a fixed data generating process that violates inequalities (1.1) for some Borel set B, the test based on T_N is consistent, i.e., the rejection probability converges to one as $N \to \infty$.

This theorem establishes asymptotic uniform validity of the proposed test procedure over $\mathcal{P}$. The second claim of the proposition concerns the power of the test at a fixed alternative, and it shows that any alternatives violating the testable implication (1.1) can be consistently rejected.

2.3 Power against $N^{-1/2}$-local Alternatives

In this section, we show that the proposed test has nontrivial power against a class of non-parametric $N^{-1/2}$-local alternatives. Let $\left\{\left(P^{[N]}, Q^{[N]}\right) \in \mathcal{P}^2 : N = 1, 2, \dots\right\}$ denote a

sequence of probability measures on $\mathcal{Y} \times \{1,0\}$ shrinking to $(P_0, Q_0) \in \mathcal{H}_0$. The next assumption defines a class of local alternatives, against which we derive power of our test.

Assumption-LA:

A sequence of true alternatives $\left\{\left(P^{[N]}, Q^{[N]}\right) \in \mathcal{P}^2 : N = 1, 2, \dots\right\}$ is represented by

$$\begin{aligned} P^{[N]} &= P_0 + N^{-1/2}\beta_1^{[N]} \text{ and} \\ Q^{[N]} &= Q_0 + N^{-1/2}\beta_0^{[N]}, \end{aligned} \tag{2.4}$$

where $(P_0, Q_0) \in \mathcal{P}^2$ is a pair of probability measures on $\mathcal{Y}\times\{1,0\}$ and $\left\{\left(\beta_1^{[N]}, \beta_0^{[N]}\right) : N = 1, 2, \dots\right\}$ is a sequence of bounded signed measures on $\mathcal{Y} \times \{1,0\}$.

(a) $(P_0, Q_0) \in \mathcal{H}_0$ and $P_0\left([y, y'], d\right) = Q_0\left([y, y'], d\right) > 0$ for some $-\infty \leq y \leq y' \leq \infty$, and $d \in \{1,0\}$.

(b) For all N, $-N^{1/2}P_0 \leq \beta_1^{[N]} < \infty$ and $-N^{1/2}Q_0 \leq \beta_0^{[N]} < \infty$ hold and $\beta_1^{[N]}(\mathcal{Y}, d) = \beta_0^{[N]}(\mathcal{Y}, d) = 0$ for $d = 1, 0$.

(c) $\beta_1^{[N]} - \beta_0^{[N]}$ converges in terms of the sup metric over Borel sets to a bounded signed measure $\Delta\beta$ as $N \to \infty$.

(d) For some $([y, y'], d)$ satisfying (a), $\Delta\beta\left([y, y'], 1\right) < 0$ and/or $\Delta\beta\left([y, y'], 0\right) > 0$ hold.

Assumption-LA (a) says that $(P_0, Q_0) \in \mathcal{H}_0$, to which $\left(P^{[N]}, Q^{[N]}\right)$ converges, has a nonempty contact set with a positive measure in terms of $P_0 = Q_0$. Assumption-LA (b) ensures $\left(P^{[N]}, Q^{[N]}\right) \in \mathcal{P}^2$ and $\Pr(D = 1|Z = 1) \geq \Pr(D = 1|Z = 0)$ for all N, and $P^{[N]}$ and $Q^{[N]}$ are in an $N^{-1/2}$-neighborhood of P_0 and Q_0 in terms of the total variation distance. Assumption-LA (c) implies that $\sqrt{N}\left(P^{[N]} - Q^{[N]}\right)\left([y, y'], d\right) \to \Delta\beta\left([y, y'], d\right)$ at every $[y, y']$ contained in the contact set of P_0 and Q_0. Accordingly, combined with Assumption-LA (d), $\left(P^{[N]}, Q^{[N]}\right)$ violates the IV-validity testable implication at some $[y, y']$ contained in the contact set for all large N.

The next theorem provides a lower bound of the power of our test along $N^{-1/2}$-local alternatives satisfying Assumption-LA.

Theorem 2.2 *Assume Condition-RG and* $\left\{\left(P^{[N]}, Q^{[N]}\right) \in \mathcal{P}^2 : N = 1, 2, \dots\right\}$ *satisfies Assumption-LA. Then,*

$$\lim_{N\to\infty} \Pr_{P^{[N]}, Q^{[N]}}\left(T_N > c_{N,1-\alpha}\right) \geq 1 - \Phi(t),$$

where $\Phi(\cdot)$ *is the cumulative distribution function of the standard normal distribution,*

$$t = \left(\frac{\sigma^2_{P_0,Q_0}\left([y,y'],1\right)}{\xi^2} \wedge 1 \right)^{-1} \left[c_{1-\alpha} - \left[\lambda(1-\lambda)\right]^{1/2} \frac{\left|\Delta\beta\left([y,y'],d\right)\right|}{\xi \vee \sigma_{P_0,Q_0}\left([y,y'],1\right)} \right],$$

$c_{1-\alpha}$ *is the limit of the bootstrap critical value of Algorithm 2.1 that is bounded and depends only on* $(\alpha, \xi, \lambda, P_0, Q_0)$*, and* $([y,y'],d)$ *is as defined in Assumption-LA (a) and (d).*

Note that the provided lower bound of the power is increasing in $\left|\Delta\beta\left([y,y'],d\right)\right|$ and it approaches one as a deviation from the null $\left|\Delta\beta\left([y,y'],d\right)\right|$ gets larger. Hence, we conclude that, for some $N^{-1/2}$-local alternatives satisfying Assumption-LA, the power is greater than the size of the test for every $\alpha \in (0, 1/2)$.

3 Extensions

3.1 A Multi-valued Discrete Instrument

The test procedure proposed above can be extended straightforwardly to a case with a multi-valued discrete instrument, $Z \in \{z_1, z_2, \ldots, z_K\}$. Let $p(z_k) = \Pr(D = 1|Z = z_k)$, and assume knowledge of the ordering of $p(z_k)$, so that without loss of generality we assume $p(z_1) \leq \cdots \leq p(z_K)$. With the multi-valued instrument, we denote the potential outcomes indexed by treatment and instrument status by $\{Y_{dz} : d = 0, 1, z = z_1, \ldots, z_K\}$, and potential selection responses by $\{D_z : z = z_1, \ldots, z_K\}$. The following assumptions guarantee that the linear two-stage least squares estimator can be interpreted as a weighted averages of the compliers average treatment effects (Imbens and Angrist (1994)).

Assumption: IV-validity for Multi-valued Discrete Z

(i) *Instrument Exclusion*: $Y_{dz_1} = Y_{dz_2} = \cdots = Y_{dz_K}$ for $d = 1, 0$, with probability one.

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

(iii) *Instrument Monotonicity: Given* $p(z_1) \leq \cdots \leq p(z_K)$, the potential selection indicators satisfy $D_{z_{k+1}} \geq D_{z_k}$ *with probability one for every* $k = 1, \ldots, (K-1)$.

Let $P(B,d|z_k) = \Pr(Y \in B, D = d|Z = z_k)$, $k = 1,\ldots,K$, and $P_{m_k}(B,d|z_k)$ be its empirical distribution based on the subsample of $Z_i = z_k$ with size m_k. The testable implication of the binary instrument case is now generalized to the following set of inequalities,

$$\begin{array}{l} P(B,1|z_1) \le P(B,1|z_2) \le \cdots \le P(B,1|z_K) \quad \text{and} \\ P(B,0|z_1) \ge P(B,0|z_2) \ge \cdots \ge P(B,0|z_K) \end{array} \tag{3.1}$$

for all Borel set B in $\mathcal{Y}$. Using the test statistic for the binary Z case to measure the violation of the inequalities across the neighboring values of Z, we can develop a statistic that jointly tests the inequalities of (3.1),

$$T_N = \max\left\{T_{N,1},\ldots,T_{N,K-1}\right\}, \tag{3.2}$$

where, for $k = 1,\ldots,(K-1)$,

$$\begin{aligned} T_{N,k} &= \left(\frac{m_k m_{k+1}}{m_k+m_{k+1}}\right)^{1/2} \max\left\{ \begin{array}{l} \sup_{-\infty\le y\le y'\le\infty}\left\{\frac{P_{m_k}([y,y'],1|z_k)-P_{m_{k+1}}([y,y'],1|z_{k+1})}{\xi_k\vee\sigma_k([y,y'],1)}\right\}, \\ \sup_{-\infty\le y\le y'\le\infty}\left\{\frac{P_{m_{k+1}}([y,y'],0|z_{k+1})-P_{m_k}([y,y'],0|z_k)}{\xi_k\vee\sigma_k([y,y'],0)}\right\} \end{array} \right\}, \\ \sigma_k^2([y,y'],d) &= \left(\frac{m_{k+1}}{m_k+m_{k+1}}\right) P_{m_k}([y,y'],d|z_k)\left(1-P_{m_k}([y,y'],d|z_k)\right) \\ &\quad + \left(\frac{m_k}{m_k+m_{k+1}}\right) P_{m_{k+1}}([y,y'],d|z_k)(1-P_{m_{k+1}}([y,y'],d|z_k)), \end{aligned}$$

and $\xi_1,\ldots,\xi_{K-1}$ are positive constants. Critical values can be obtained by applying a resampling algorithm of the previous section to each $T_{N,k}$ simultaneously.

Algorithm 3.1:

(i) Let $H_{N,k}(\cdot) = \left(\frac{m_{k+1}}{m_k+m_{k+1}}\right) P_{m_{k+1}}(\cdot|z_{k+1}) + \left(\frac{m_k}{m_k+m_{k+1}}\right) P_{m_k}(\cdot|z_k)$ be the pooled empirical measures that pool the sample of $Z_i = z_{k+1}$ and that of $Z_i = z_k$. *Sample* (Y_i^*, D_i^*), $i = 1,\ldots,m_{k+1}$ *randomly with replacement from* $H_{N,k}$ *and construct the bootstrap empirical distribution* $P^*_{m_{k+1}}(\cdot|z_{k+1})$. *Similarly, sample* (Y_j^*, D_j^*), $j = 1,\ldots,m_k$ *randomly with replacement from* $H_{N,k}$ *and construct the bootstrap empirical distribution* $P^*_{m_k}(\cdot|z_k)$.

(ii) Apply step 1 for every $k = 1,\ldots,(K-1)$, and obtain $(K-1)$ pairs of the resampled empirical measures, $(P^*_{m_1}, P^*_{m_2})$, $(P^*_{m_2}, P^*_{m_3}),\ldots,(P^*_{m_{K-1}}, P^*_{m_K})$. Define, for

$k = 1, \ldots, (K-1)$,

$$\begin{aligned}
\sigma_k^{*2}([y,y'],d) &= \left(\frac{m_{k+1}}{m_k+m_{k+1}}\right) P^*_{m_k}([y,y'],d|z_k)\left(1-P^*_{m_k}([y,y'],d|z_k)\right) \\
&\quad + \left(\frac{m_k}{m_k+m_{k+1}}\right) P^*_{m_{k+1}}([y,y'],d|z_{k+1})(1-P^*_{m_{k+1}}([y,y'],d|z_{k+1})), \\
T^*_{N,k} &= \left(\frac{m_k m_{k+1}}{m_k+m_{k+1}}\right)^{1/2} \\
&\quad \times \max\left\{ \begin{array}{c} \sup_{-\infty\le y\le y'\le\infty}\left\{\frac{P^*_{m_k}([y,y'],1|z_k)-P^*_{m_{k+1}}([y,y'],1|z_{k+1})}{\xi_k\vee\sigma^*_k([y,y'],1)}\right\}, \\ \sup_{-\infty\le y\le y'\le\infty}\left\{\frac{P^*_{m_{k+1}}([y,y'],0|z_{k+1})-P^*_{m_k}([y,y'],0|z_k)}{\xi_k\vee\sigma^*_k([y,y'],0)}\right\} \end{array} \right\},
\end{aligned}$$

where ξ_k, $k = 1, \ldots, (K-1)$, are positive constants. *The bootstrap statistic* T^*_N is computed accordingly by $T^*_N = \max\left\{T^*_{N,1}, \ldots, T^*_{N,K-1}\right\}$.

(iii) *Iterate Step 1 -3 many times, get the empirical distribution of* T^*_N, and *obtain a bootstrap critical value* $c_{N,1-\alpha}$ *from its empirical* $(1-\alpha)$*-th quantile .*

(iv) *Reject the null hypothesis (3.1) if* $T_N > c_{N,1-\alpha}$. *The bootstrap p-value is obtained by the proportion of* T^*_N*'s greater than* T_N.

3.2 Conditioning Covariates

Empirical studies commonly use observable conditioning covariates in the context of instrumental variable methods. One of the major motivations for introducing them is to control for potential confounders that invalidate the random assignment assumption. This section briefly discusses how to extend IV-validity test proposed above to the settings with conditioning covariates, $X \in \mathbb{X} \subset \mathbb{R}^{d_x}$, used for this purpose.

IV-validity to be tested in this case consists of the joint restriction of instrument exclusion, instrument monotonicity, and the conditional version of the instrument random assignment assumption, $(Y_{11}, Y_{10}, Y_{01}, Y_{00}, D_1, D_0) \perp Z|X$. These three assumptions combined with the first stage rank condition, $\Pr(D=1|Z=1,X) \neq \Pr(D=1|Z=0,X)$ for some X, guarantee that the linear two stage least squares with a function of (Z,X) used as an instrument (e.g. $\Pr(D=1|Z,X)$) estimates a certain weighted average of the complier's conditional causal effects $E(Y_1 - Y_0|D_1 > D_0, X)$ (Heckman and Vytlacil (2005)). Moreover, under the same set of assumptions, the semiparametric IV estimator developed by Abadie (2003) consistently estimates the unconditional complier's causal effect $E(Y_1 - Y_0|D_1 > D_0)$.

A testable implication with the largest screening power in the sense similar to Proposition 1.1 (i) is given by the conditional version of the inequalities (1.1), i.e., for every Borel set $B \subset \mathcal{Y}$ and $X \in \mathbb{X}$,

$$\begin{aligned} &\Pr(Y \in B, D = 1|Z = 1, X) - \Pr(Y \in B, D = 1|Z = 0, X) \geq 0, \\ &\Pr(Y \in B, D = 0|Z = 0, X) - \Pr(Y \in B, D = 1|Z = 1, X) \geq 0. \end{aligned} \tag{3.3}$$

As shown in Kitagawa (2015, Lemma B.8), the use of Theorem 3.1 of Abadie (2003) and the instrument function argument for conditional moment inequalities as given in Andrews and Shi (2013) and Khan and Tamer (2009) enable us to reduce (3.3) to the following unconditional moment inequalities without loss of any information,[11]

$$\begin{aligned} E\left[\kappa_1\left(D, Z, X\right) g(Y, X)\right] &\geq 0, \\ E\left[\kappa_0\left(D, Z, X\right) g(Y, X)\right] &\geq 0, \quad \text{for all } g(\cdot,\cdot) \in \mathcal{G} \end{aligned} \tag{3.4}$$

where

$$\begin{aligned} \kappa_1\left(D, Z, X\right) &= D\frac{Z - \Pr(Z = 1|X)}{\Pr(Z = 0|X)\Pr(Z = 1|X)}, \\ \kappa_0\left(D, Z, X\right) &= (1 - D)\frac{(1 - Z) - \Pr(Z = 0|X)}{\Pr(Z = 0|X)\Pr(Z = 1|X)}, \end{aligned}$$

and $\mathcal{G}$ is the class of indicator functions for boxes in $\mathcal{Y} \times \mathcal{X}$,

$$\mathcal{G} = \left\{ \begin{array}{c} 1\{(Y, X) \in C\} : C = [y, y'] \times [x_1, x_1'] \times \cdots \times \left[x_{d_x}, x_{d_x}'\right], \ -\infty \leq y \leq y' \leq \infty, \\ -\infty \leq x_l \leq x_l' \leq \infty, \quad l = 1, \ldots, d_x. \end{array} \right\}. \tag{3.5}$$

Accordingly, a variance-weighted KS statistic to infer (3.4) can be proposed as

$$T_N = \sqrt{N} \max \left\{ \begin{array}{c} \sup_{g \in \mathcal{G}} \frac{-E_N[\hat{\kappa}_1(D,Z,X)g(Y,X)]}{\xi \vee \hat{\sigma}_1(g)}, \\ \sup_{g \in \mathcal{G}} \frac{-E_N[\hat{\kappa}_0(D,Z,X)g(Y,X)]}{\xi \vee \hat{\sigma}_0(g)} \end{array} \right\},$$

[11] If the random assignment assumption is strengthened to $(Y_{11}, Y_{10}, Y_{01}, Y_{00}, D_1, D_0, X) \perp Z$, then it can be shown that the moment conditions of (3.4) are reduced to

$$\begin{aligned} &\Pr((Y, X) \in C, D = 1|Z = 1) - \Pr((Y, X) \in C, D = 1|Z = 0) \geq 0, \\ &\Pr((Y, X) \in C, D = 0|Z = 0) - \Pr((Y, X) \in C, D = 0|Z = 1) \geq 0, \end{aligned}$$

for any box C in $\mathcal{Y} \times \mathcal{X}$. As a result, the test procedure for no-covariate case can be extended straightforwardly to this case.

where $\hat{\kappa}_d$ is an estimate of κ_d with estimated $\Pr(Z=1|X)$ plugged in, $E_N(\cdot)$ is the sample average, and $\hat{\sigma}_d^2(g)$ is the sample variance of $\hat{\kappa}_d(D_i, Z_i, X_i)\, g(Y_i, X_i)$. Treating $\hat{\kappa}_d$ as given (estimated from the original sample), we obtain the critical values by bootstrapping the supremum statistic of the recentered moments,

$$T_N^* = \sqrt{N} \max \left\{ \begin{array}{c} \sup_{g\in\mathcal{G}} \frac{-\left[E_N^*[\hat{\kappa}_1(D,Z,X)g(Y,X)] - E_N[\hat{\kappa}_1(D,Z,X)g(Y,X)]\right]}{\xi \vee \hat{\sigma}_1^*(g)}, \\ \sup_{g\in\mathcal{G}} \frac{-\left[E_N^*[\hat{\kappa}_0(D,Z,X)g(Y,X)] - E_N[\hat{\kappa}_0(D,Z,X)g(Y,X)]\right]}{\xi \vee \hat{\sigma}_0^*(g)} \end{array} \right\},$$

where $E_N^*(\cdot)$ is the sample average based on a bootstrap sample that is obtained by resampling (Y, D, Z, X) from the original sample, and $\hat{\sigma}_d^{*2}(g)$ is the variance estimate based on the bootstrap sample.[12]

In terms of practical implementation, a couple of issues deserve attention. First, in the presence of many covariates, computation of the statistic involves an optimization over a large class of indicator functions. This raises a computational challenge in implementing the test. Second, validity of the test relies on consistent estimation of $\Pr(Z=1|X)$. Hence, if a parametric estimation for $\Pr(Z=1|X)$ is used to implement the test, a misspecified functional form in the estimation of $\Pr(Z=1|X)$ can lead to an erroneous conclusion.

4 Empirical Applications

We illustrate a use of our test using the two data sets mentioned in Introduction.

4.1 Draft Lottery Data

The draft lottery data consist of a sample of 11,637 white men, born between 1950 and 1953 extracted from March Current Population Surveys of 1979 and 1981-1985. This data set is a subsample of the sample used in Angrist and Krueger (1992, 1995). Following Abadie (2002), we define a binary draft eligibility instrument by a dummy variable indicating whether or not one's lottery number is smaller than or equal to 100. See Angrist (1990) for detailed description of the Vietnam era draft lottery. We apply our test to two outcome measures, annual labor earnings and weekly wages, which are measured in terms of 1978 dollars using

[12] We leave for future research a formal investigation on influences of estimation errors in $\hat{\kappa}_d$ to the performance of our test.

the CPI. The measure of weekly earnings is imputed by the annual labor earnings divided by the weeks worked. The treatment is whether one has a Vietnam veteran status. Since the draft lottery numbers are randomly assigned based on one's birthdate, it is reasonable to believe that the constructed instrument is independent of any individual characteristics. It is hard to believe existence of defiers in the current context even though the sampling design does not exclude the possibility of having them. A less credible assumption would be instrument exclusion. For instance, the draft lottery can directly affect control outcomes for some never-takers if those who were drafted change their career choice, school years, or migration choice for the purpose of escaping from the military service.

Table I shows the result of our test. We present the bootstrap p-values of our test for several different specifications of the trimming constant. All of them are close to or exactly equal to one. Hence, we do not reject validity of the draft lottery instrument from the data.

4.2 Returns to Education: Proximity to College Data

The Card data is based on National Longitudinal Survey of Young Men (NLSYM) that began in 1966 with 14-24 years old men and continued with follow-up surveys through 1981. Based on the respondents' county of residence at 1966, the Card data provides the presence of a 4-year college in the local labor market. The observations of years of education and wages were based on the follow-ups' educational attainment and wages reported in the interview in 1976.

Proximity to college was used as an instrument, because the presence of a nearby college reduces the cost of college education by allowing students to live at their home, while one's unobservable ability is presumably independent of student's residence during their teenage years. Compliers in this context can be considered as those who grew up in relatively low-income families and who were not able to go to college without living with their parents. Being different from the original Card's study, we treat the educational level as a binary treatment, with years of education greater than or equal to 16 years, that is, the treatment can be considered as a four year college degree.

We specify the measure of outcome to be the logarithm of weekly earnings. In the first specification, we do not control for any demographic covariates. This raises a concern regarding the violation of random assignment assumption. For instance, one's region of residence, or whether they were born in the standard metropolitan area or rural area. may

well be dependent on one's wage levels and the proximity to colleges if the urban areas are more likely to have colleges and higher wage levels compared to the rural areas.

Our test procedure yields zero p-values for each choice of trimming constant. This provides an empirical evidence that without controlling for any covariates, college proximity is not a valid instrument.

Table I: Test Results of the Empirical Applications

Bootstrap iterations 500

data	Draft lottery data						Proximity to college data					
sample size (m,n)	(3234,8403)						(2053,957)					
$\Pr(D=1\|Z=1)$, $\Pr(D=1\|Z=0)$	0.29, 0.18						0.29, 0.22					
Y	annual earnings			weekly wages			weekly wages			weekly wages		
	No Covariate			No Covariate			No Covariate			With Covariates*		
trimming constant ξ	0.07	0.3	1	0.07	0.3	1	0.07	0.3	1	0.07	0.3	1
Bootstrap test, p-value	0.93	1.00	1.00	1.00	1.00	1.00	0.00	0.00	0.00	0.89	0.71	0.91

* five dummy variables indicating (i) residence in a standard metropolitan area (SMSA) in 1976, (ii) residence in SMSA in 1966, (iii) race is black or not, (iv) residence in southern states in 1976, and (v) residence in southern states in 1966.

The original study of Card (1993) indeed emphasized the importance of controlling for regions, residence in the urban area, race, job experience, and parent's education, and he included them in his specification of the two stage least square estimation. In our second specification, we control for the covariates listed at the bottom of Table I, which are all binary variables. We estimate $\Pr(Z=1|X)$ using a linear probability regression with these five dummy variables. The class of indicator functions $\mathcal{G}$ we use is

$$\mathcal{G} = \left\{ \begin{array}{c} 1\{(Y,X)\in C\} : C = [y_q, y_{q'}] \times \{x_1\} \times \cdots \times \{x_{d_x}\}, \\ y_q \text{ is the empirical } q\text{-th quantile of } Y, \\ q,\, q' \in \{0, 0.05, \ldots, 0.95, 1\},\, q < q' \\ x_l \in \{0,1\}, \quad l = 1, \ldots, d_x. \end{array} \right\}$$

With these covariates, the p-values turn out to be large. We therefore conclude that we do not reject validity of the college proximity instrument once these covariates are controlled for.

References

[1] Abadie, A. (2002): "Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models," *Journal of the American Statistical Association*, 97, 284-292.

[2] Abadie, A. (2003): "Semiparametric Instrumental Variable Estimation of Treatment Response Models," *Journal of Econometrics*, 113, 231 - 263.

[3] Abadie, A., J. D. Angrist, and G. W. Imbens. (2002): "Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings," *Econometrica*, 70, 91-117.

[4] Andrews, D.W.K. and P. Jia Barwick (2012): "Inference for Parameters Defined by Moment Inequalities: A Recommended Moment Selection Procedure," *Econometrica*, 80, 2805-2826.

[5] Andrews, D.W.K. and X. Shi (2013): "Inference Based on Conditional Moment Inequalities," Econometrica, 81, 609-666.

[6] Andrews, D.W.K. and G. Soares (2010): "Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection," *Econometrica*, 78, 119-157.

[7] Angrist, J. D. (1990): "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records," *The American Economic Review*, 80, 313-336.

[8] Angrist, J.D. and G.W. Imbens (1995): "Two-stage Least Squares Estimation of Average Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91, 444-455.

[9] Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996): "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91, 444-455.

[10] Angrist, J.D. and A. B. Krueger (1992): "Estimating the Payoff to Schooling Using the Vietnam-Era Draft Lottery," *NBER Working Paper Series*, No. 4067.

[11] Angrist, J.D. and A. B. Krueger (1995): "Split-Sample Instrumental Variables Estimates of the Return to Schooling," *Journal of Business and Economic Statistics*, Vol. 13, No.2, pp. 225-235.

[12] Armstrong, T.B. and H.P. Chan (2013): "Multiscale Adaptive Inference on Conditional Moment Inequalities," *Cowles Foundation Discussion Paper*, No. 1885. Yale University.

[13] Armstrong, T.B. (2014): "Weighted KS Statistics for Inference on Conditional Moment inequalities," *Journal of Econometrics*, 181, 92-116.

[14] Balke, A. and J. Pearl (1997): "Bounds on Treatment Effects from Studies with Imperfect Compliance," *Journal of the American Statistical Association*, 92, 1171-1176.

[15] Barrett, G.F. and S.G. Donald (2003): "Consistent Tests for Stochastic Dominance," *Econometrica* 71, 71-104.

[16] Barua, R. and K. Lang (2009): "School Entry, Educational Attainment and Quarter of Birth: A Cautionary Tale of LATE." *NBER Working Paper* 15236, National Bureau of Economic Research.

[17] Breusch, T.S. (1986): "Hypothesis Testing in Unidentified Models," *Review of Economic Studies*, 53, 4, 635-651.

[18] Card, D. (1993): "Using Geographical Variation in College Proximity to Estimate the Returns to Schooling", National Bureau of Economic Research Working Paper No. 4, 483.

[19] Chernozhukov, V., S. Lee, and A.M. Rosen (2013): "Intersection Bounds: Estimation and Inference," *Econometrica*, 81, 667–737.

[20] Chetverikov, D. (2012): "Adaptive Test of Conditional Moment Inequalities," *unpublished manuscript*, UCLA.

[21] Donald, S. and Y.-C. Hsu (forthcoming): "Improving the Power of Tests of Stochastic Dominance," *forthcoming in Econometric Reviews.*

[22] Fiorini, M., K. Stevens, M. Taylor, and B. Edwards (2013): "Monotonically Hopeless? Monotonicity in IV and Fuzzy RD Designs," *unpublished manuscript*, University of Technology Sydney, University of Sydney, and Australian Institute of Family Studies.

[23] Heckman, J. J. and E. Vytlacil (2005): "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica* 73, 669-738.

[24] Horváth, L., P. Kokoszka, and R. Zitikis (2006): "Testing for Stochastic Dominance Using the Weighted McFadden-type statistic," *Journal of Econometrics*, 133, 191-205.

[25] Huber, M. and G. Mellace (2013) "Testing Instrument Validity for LATE Identification based on Inequality Moment Constraints," *unpublished manuscript*, University of Sankt Gallen.

[26] Imbens, G.W. (forthcoming) "Instrumental Variables: An Econometrician's Perspective," *forthcoming in Statistical Science*.

[27] Imbens, G. W. and J. D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467-475.

[28] Imbens, G. W. and D. B. Rubin (1997): "Estimating Outcome Distributions for Compliers in Instrumental Variable Models," *Review of Economic Studies*, 64, 555-574.

[29] Khan, S. and E. Tamer (2009): "Inference on Endogenously Censored Regression Models Using Conditional Moment Inequalities," *Journal of Econometrics*, 152, 104 - 119.

[30] Kitagawa, T. (2015): "A Test for Instrument Validity, Supplementary Material: Proofs and Monte Carlo Experiments," *Online Material in Econometrica*.

[31] Lee, SJ. (2014): "On Variance Estimation for 2SLS When Instruments Identify Different LATEs," *unpublished manuscript*, University of New South Wales.

[32] Lee, S., K. Song, and Y. Whang (2011): "Testing Functional Inequalities," *cemmap working paper* 12/11, University College London.

[33] Linton, O., E. Maasoumi, and Y. Whang (2005): "Consistent Testing for Stochastic Dominance under General Sampling Schemes," *Review of Economic Studies*, 72, 735-765.

[34] Linton, O., K. Song, and Y. Whang (2010): "An Improved Bootstrap Test of Stochastic Dominance," *Journal of Econometrics*, 154, 186-202.

[35] Machado, C., A. M. Shaikh, and E. J. Vytlacil (2013): "Instrumental Variables and the Sign of the Average Treatment Effect," *unpublished manuscript*, Getulio Vargas Foundation, University of Chicago, and New York University.

[36] Mourifié, I. and Y. Wan (2014): "Testing LATE Assumptions," *unpublished manuscript*, University of Toronto.

[37] Romano, J. P. (1988): "A Bootstrap Revival of Some Nonparametric Distance Tests." *Journal of American Statistical Association*, 83, 698-708.