Xavier de Luna* and Per Johansson

Testing for the Unconfoundedness Assumption Using an Instrumental Assumption

Abstract: The identification of average causal effects of a treatment in observational studies is typically based either on the unconfoundedness assumption (exogeneity of the treatment) or on the availability of an instrument. When available, instruments may also be used to test for the unconfoundedness assumption. In this paper, we present a set of assumptions on an instrumental variable which allows us to test for the unconfoundedness assumption, although they do not necessarily yield nonparametric identification of an average causal effect. We propose a test for the unconfoundedness assumption based on the instrumental assumptions introduced and give conditions under which the test has power. We perform a simulation study and apply the results to a case study where the interest lies in evaluating the effect of job practice on employment.

Keywords: average treatment effect, job practice, nonparametric identification

DOI 10.1515/jci-2013-0011

1 Introduction

Identification of the causal effect of a treatment T on an outcome Y in observational studies is typically based either on the unconfoundedness assumption (also called selection on observables, exogeneity, ignorability, see, e.g. de Luna and Johansson [1], Imbens and Wooldridge [2], Pearl [3]) or on the availability of an instrument. The unconfoundedness assumption says loosely that all the variables affecting both the treatment T and the outcome Y are observed (we call them covariates) and can be controlled for. An instrument is usually defined as a variable affecting the treatment T , and such that it is related to the outcome Y only through T (and possibly the observed covariates). When available, instruments can be used to identify causal effects in parametric situations. Nonparametric identification is also possible with the help of instruments, and Angrist et al. [4] develop a theory for the nonparametric identification and estimation of local average causal effects. Abadie [5] and Frölich [6] extended these results to the situation where the observed covariates are related to the instrument. Note also that nonparametric identification can be obtained with the related concept of (fuzzy) regression discontinuity designs; see Hahn et al. [7], Battistin and Retore [8], Dias et al. [9] and Lee [10, Sec. 5.5.3]. When a causal effect is identified, a test of the unconfoundedness assumption may be devised by comparing the estimates of the causal effects obtained both under the unconfoundedness assumption and using the instrument (classical Durbin–Wu–Hausman (DWH) test in a parametric setting). This was recently used by Donald et al. [11] to propose a test of the unconfoundedness assumption in a nonparametric framework.

In this paper, we introduce general instrumental conditions under which it is possible to test for the unconfoundedness or exogeneity assumption. The instrumental assumptions are general and, for instance,

*Corresponding author: **Xavier de Luna**, Department of Statistics, Umeå School of Business and Economics, Umeå University, SE-90187 Umeå, Sweden, E-mail: xavier.deluna@stat.umu.se

Per Johansson, Department of Economics, Uppsala University, Uppsala, Sweden; Institute for Evaluation of Labour Market and Education Policy, Uppsala, Sweden, E-mail: Per.Johansson@ifau.uu.se

they do not necessarily yield identification of a causal effect when the unconfoundedness assumption does not hold. Indeed, to obtain the nonparametric identification of a local average causal effects stronger (and untestable) assumptions must be made on the instrument, see, e.g. Imbens and Angrist [12], Angrist et al. [4], Angrist and Fernandez-Val [13], Donald et al. [11] and Guo et al. [14]. In particular, these papers use a monotonicity assumption saying that the instrument must affect the treatment in a monotone fashion, as well as do not allow for unobserved heterogeneity to affect both the instrument and the treatment. Based on our general instrumental conditions we can propose a statistic to test the unconfoundedness assumption. The proposed test is related to the use of two control groups to test the unconfoundedness assumption, an idea previously used, e.g. in Rosenbaum [15], de Luna and Johansson [1] and Dias et al. [9]. Rosenbaum [15] was probably first to formalize the idea that two control groups provide information on the unconfoundedness assumption and described actual observational studies where different control groups were available. One of our contributions in this context is the introduction of general assumptions under which an observed variable can be used to split an available control group in order to test the unconfoundedness assumption nonparametrically. However, the test statistic we eventually propose does not actually require the split to be done.

In Section 4, we present a motivating example where Swedish register data are used to study the causal effect of job practice (JP) on employment. We have access to a rich set of background characteristics on unemployed individuals, although the question remains whether the effect of JP on employment is confounded by unobserved heterogeneity. In this study, unemployed have access to JP through their participation into a labor market program. During 1998 there were two such labor market programs available in Sweden offering JP with different probabilities. Because we know that the two programs differ mainly only with respect to their propensity to offer JP, the participation into the two programs may be assumed to affect employment differently only through JP. We, thus, argue that program participation fulfills our instrumental conditions. In contrast with usual instrumental assumptions this allows potential unobserved heterogeneity in the program and JP assignment to be correlated. We apply the introduced test to check whether the estimated effect of JP on employment is biased due to unobservables affecting both JP and employment.

Before treating this motivating example in more details in Section 4, Section 2 presents the model, introduces instrumental assumptions and develops the theoretical results which then allow us to introduce a test of the unconfoundedness assumption. Section 3 presents a simulation study of the finite sample properties of the proposed test. In particular, one of the designs used illustrates the situation where the monotonicity assumption mentioned above does not hold. The paper is concluded in Section 5.

2 Theory and method

2.1 Model

We use the Neyman–Rubin model [16, 17] for causal inference when the interest lies in the causal effect of a binary treatment T , taking values in $\mathcal{T} = \{0, 1\}$, on an outcome. Let us thus define $Y(t)$, $t \in \mathcal{T}$, called potential outcomes. The latter are interpreted as the outcomes resulting from the assignment $T = t$, $t \in \mathcal{T}$, respectively. We then observe $Y = TY(1) + (T - 1)Y(0)$. Let us also assume that we observe a set of variables which are not affected by the treatment assignment. We will need to distinguish in particular $\mathbf{X}$ and $\mathbf{Z}$ two vectors of such variables, the latter of dimension one.

For $t \in \mathcal{T}$, we consider $(\mathbf{X}, \mathbf{Z}, T, Y(t))$ as a random vector variable with a given joint distribution, from which a random sample is drawn. Population parameters that are often of interest in this context are the average causal effect $\theta = E(Y(1) - Y(0))$ and the average causal effect on the treated $\theta^t = E(Y(1) - Y(0) | T = 1)$ or on the non-treated $\theta^{nt} = E(Y(1) - Y(0) | T = 0)$.

In observational studies, where the treatment assignment T is not randomized, an identifying assumption (e.g. Rosenbaum and Rubin [18]; Imbens [19]) for the average causal effect is the following.

(A.1) For $t \in \mathcal{T}$,

$$\begin{aligned} T &\perp\!\!\!\perp Y(t)|\mathbf{X} && \text{(unconfoundedness)}, \\ \Pr(T = t|\mathbf{X}) &> 0 && \text{(common support)}. \end{aligned}$$

The common support assumption can be investigated by looking at the data. The unconfoundedness assumption may be considered as realistic in situations where the set of characteristics $\mathbf{X}$ is rich enough, and when there is subject-matter theory to support the assumption.

2.2 Instrumental assumptions, test and power

Let us now consider situations where the variable Z takes values in $\mathcal{T}$ (if not, it may be made dichotomous using a threshold) and fulfills the following assumption.

(A.2) For $t \in \mathcal{T}$,

$$\begin{aligned} Z &\perp\!\!\!\perp Y(t)|\mathbf{X}, \\ 0 &< \Pr(Z = 1|\mathbf{X}) < 1. \end{aligned}$$

Assumption (A.2) prohibits (a) a direct effect from Z to $Y(t)$, i.e. an effect not going through T and (b) unobserved variables affecting both Z and $Y(t)$. On the other hand, (A.2) allows unobserved variables to affect both Z and T which is typically prohibited by usual instrumental assumptions [4–6]. Note that when assuming (A.2) in the sequel, Z and $Y(t)$ may also be independent conditional on a subset of $\mathbf{X}$, and, e.g. Z may be randomized as discussed after Proposition 1. We also need the following regularity condition.

(A.3) If (A.1) and (A.2) hold, then $T \perp\!\!\!\perp Y(t)|Z, \mathbf{X}$, for $t \in \mathcal{T}$ respectively.

Assumption (A.3) is a regularity condition and is violated only in specific situations, of which Example 1 is typical.

Example 1 Let us assume that the vector $(Z^*, T^*, Y(0), U, V)$ has joint normal distribution, where U and V are two unobserved covariates and the set of observed covariates $\mathbf{X}$ is empty. Assume now that the following model generates the data:

$$\begin{aligned} Z^* &= \psi_0 + \psi_1 U + \psi_2 V + \varepsilon_Z, \\ T^* &= v_0 + v_1 V + \varepsilon_T, \\ Y(0) &= \xi_1 Z^* + \xi_2 U + \varepsilon_Y. \end{aligned} \tag{1}$$

Figure 1 Graph illustrating model (1) in Example 1

where $U, V, \varepsilon_Z, \varepsilon_T$ and ε_Y are jointly normal and independently distributed. Let $Z = \mathcal{I}(Z^* > 0)$ and $T = \mathcal{I}(T^* > 0)$, where $\mathcal{I}(\cdot)$ is the indicator function. Figure 1 gives a graphical representation of the model, where $\varepsilon_Z, \varepsilon_T$ and ε_Y are omitted. We can write the conditional expectations

$$E(Y(0)|Z^*, U) = \xi_1 Z^* + \xi_2 U,$$

$$E(U|Z^*) = \gamma Z^*,$$

where γ is function of the parameters in (1).

In Example 1, (A.1) and (A.2) will typically be violated, unless we assume that $\xi_1 = -\xi_2\gamma$, in which case $Z^* \perp\!\!\!\perp Y(0)$ by joint normality, and thereby $Z \perp\!\!\!\perp Y(0)$ and $T \perp\!\!\!\perp Y(0)$. The constrained parametrization $\xi_1 = -\xi_2\gamma$ yields thus an example where (A.3) is violated since (A.1) and (A.2) hold while one can check that $T \perp\!\!\!\perp Y(0)|Z$ does not necessary hold.

This type of example is called unstable [3, Sec. 2.4] in the sense that (A.1 and A.2) will cease to hold as soon as the parameter values do not fulfill the constraint $\xi_1 = -\xi_2\gamma$. Using directed acyclic graphs,¹ it can be shown that assumption (A.3) holds as soon as the distribution is stable, where, e.g. a distribution $P(\psi)$ parametrized with a parameter vector ψ is said stable if no independence can be destroyed by varying the parameter ψ ; see Pearl [3, Sec. 2.4] for a formal general definition. Note here that (A.3) does not imply any parametrized functional form.

Proposition 1 Assume (A.1)–(A.3), then

$$Z \perp\!\!\!\perp Y(t)|T, \mathbf{X}, \quad t \in \mathcal{T}. \quad (2)$$

Proof. By assumption (A.1) and (A.2) hold. Then, for $t \in \mathcal{T}$,

$$\begin{aligned} (A.1) \text{ and } (A.2) &\Rightarrow T \perp\!\!\!\perp Y(t)|Z, \mathbf{X} \text{ and } Z \perp\!\!\!\perp Y(t)|\mathbf{X} \\ &\Rightarrow (T, Z) \perp\!\!\!\perp Y(t)|\mathbf{X} \\ &\Rightarrow Z \perp\!\!\!\perp Y(t)|T, \mathbf{X}. \end{aligned}$$

The first implication by assumption (A.3), the two other by the properties of conditional independence relations, see Dawid [21], Lauritzen [22, Sec. 3.1] and Pearl [3, Sec. 1.1.5]. ■

The conditional independence statement obtained in Proposition 1 is testable from the data when conditioning on $T = t$ (see next section). Finding evidence in the data against (2) yields evidence against the assumptions of the proposition. Thus, evidence against (2) can be interpreted as evidence against the unconfoundedness assumption (A.1) if (A.2) is known to hold from subject-matter considerations – (A.3) being a regularity condition. One application is a random experiment (where Z is a random assignment to a treatment) with restricted compliance T [4, 12]. Another example of application is treated in detail in Section 4. Note that while identification of the causal effect of T on Y may follow from (A.2) with linear models, see, e.g. Pearl [3, p. 248], this is not true in general, and stronger assumptions are needed to obtain nonparametric identification of a causal effect such as, e.g. a local average treatment effect [4–6]. In particular, our result does not rely on two assumptions typically made to obtain such identification; that the instrument must affect the treatment in a monotone fashion and that no unobserved heterogeneity is allowed to affect both the instrument and the treatment.

For a test based on (2) to have power against (A.1) we further need to have that Z and T are dependent conditional on $\mathbf{X}$. This is typically assumed for instrumental variables to be useful for identification. Examples of situations (expressed with directed acyclic graphs; see Footnote 1) for a test that would be

¹ Directed acyclic graphs, e.g. Figure 1, together with a stable (also called faithful) distribution for the variables are used to describe conditional independence relations between variables; see Lauritzen [22] for a general account on graphical models and de Luna et al. [20] for their use together with potential outcomes.

Figure 2 Four directed acyclic graphs together with a respective stable joint distribution for the variables included: Only cases (a)–(c) are such that a test based on (2) may have power, i.e. if (A.1) does not hold, e.g. through the introduction of a variable V with arrows pointing toward T and $Y(t)$, then $Y(t) \perp\!\!\!\perp Z|T, \mathbf{X}$ would not hold either

based on (2) to have power against (A.1) are given in Figure 2, panels (a)–(c), while panel (d) shows a case where such a test would not have power. A caveat here is that (2) can be tested only when conditioning on $T = t$. This has no practical consequence if the test rejects this null hypothesis. On the other hand, in cases where (2) is not rejected for $T = t$, we have no information on whether it is violated for $T = 1 - t$. In independent and related work, Guo et al. [14, eqs (3) and (4)] give an example where (2) holds for $T = t$ although not for $T = 1 - t$, and yet a specific causal effect is identified without the help of Z when the earlier mentioned monotonicity assumption holds.

2.3 Method

Different strategies may be adopted to test two null hypotheses given by Proposition 1, i.e.

$$H_0^a : Z \perp\!\!\!\perp Y(0)|T = 0, \mathbf{X},$$

$$H_0^b : Z \perp\!\!\!\perp Y(1)|T = 1, \mathbf{X}.$$

Note that for θ^t , (A.1)–(A.3) need to hold only for $t = 0$ and, thus, only H_0^a is to be tested. Similarly, H_0^b is relevant when θ^{nt} is of interest, while both null hypotheses are relevant for θ . In this paper we propose a testing strategy² based on the fact that under H_0^a and H_0^b we have $\delta_0(\mathbf{X}) = 0$ and $\delta_1(\mathbf{X}) = 0$, for all $\mathbf{X}$, respectively, where

$$\delta_0(\mathbf{X}) = E(Y|T = 0, \mathbf{X}, Z = 1) - E(Y|T = 0, \mathbf{X}, Z = 0),$$

$$\delta_1(\mathbf{X}) = E(Y|T = 1, \mathbf{X}, Z = 1) - E(Y|T = 1, \mathbf{X}, Z = 0).$$

Given a random sample of n individuals indexed by i , $i = 1, \dots, n$, we consider a nonparametric estimator for $\delta_j = E(\delta_j(\mathbf{X}_i))$, $j = 0, 1$,

$$\hat{\delta}_j = \frac{1}{N_{j1}} \sum_{i:T_i=j, Z_i=1} (Y_i - \hat{Y}_{ji}) + \frac{1}{N_{j0}} \sum_{i:T_i=j, Z_i=0} (\tilde{Y}_{ji} - Y_i),$$

² One related strategy could be to use the concept of two independent control groups [15]. Under H_0^a we can use Z to obtain two independent control groups (one defined by $Z = 1$ and one by $Z = 0$) for estimating θ , yielding $\hat{\theta}^{z=0}$ and $\hat{\theta}^{z=1}$, respectively. Under H_0^b the difference $\hat{\theta}^{z=0} - \hat{\theta}^{z=1}$ has expectation zero and this makes the basis for a test statistic. However, since we need to compute two nonparametric estimators of θ , the resulting statistic has poor finite sample properties, for instance, when the covariates have different support in the two control groups created. This has been confirmed in simulation experiments not presented here.

where $N_{jk} = \text{card}(\{i : T_i = j, Z_i = k\})$, $k = 0, 1$, with $\text{card}(A)$ denoting the cardinality of the set A , and $\hat{Y}_{ji}$ and $\tilde{Y}_{ji}$ are nonparametric estimators of $E(Y_i | T_i = j, \mathbf{X}_i, Z_i = 0)$ and $E(Y_i | T_i = j, \mathbf{X}_i, Z_i = 1)$, respectively. The two latter estimates may be obtained by nearest neighbor matching, or any other smoothing technique. Since $\delta_0 = 0$ and $\delta_1 = 0$, respectively, under H_0^a and H_0^b , the test statistics

$$C_0 = \frac{\hat{\delta}_0}{s_0} \quad \text{and} \quad C_1 = \frac{\hat{\delta}_1}{s_1} \quad (3)$$

will then, under the necessary regularity conditions, be asymptotically normally distributed with mean zero and variance one, where s_j is the standard error of $\hat{\delta}_j$, for $j = 0, 1$. For instance, if nearest neighbor matching estimators are used, then the asymptotic theory and in particular s_j can be found in Abadie and Imbens [23]. A subsampling estimator of s_j is also available in this case in de Luna et al. [24]. As noted above, when θ is of interest, then both hypotheses H_0^a and H_0^b are relevant and higher power may be obtained by considering the joint statistic

$$C = C_0^2 + C_1^2,$$

which is asymptotically χ_2^2 distributed, since C_0 and C_1 are independent.

We should note here that the statistics above are testing conditional mean independence, which is relevant when average causal effects are targeted. Alternatively, one may wish to use tests of conditional independence statements based on all the moments of the underlying distribution [25], thereby making the methods relevant when quantile or distributional causal effects are of interest.

3 Monte Carlo study

We use a Monte Carlo study to investigate the finite sample properties (empirical size and power) of the test C_0 in (3), where K -nearest neighbor matching is used as nonparametric estimator of $\hat{Y}_i(0)$ and $\tilde{Y}_i(0)$, together with the Abadie and Imbens [23] variance estimator. As noted above, in situations where θ is of interest and (A.1)–(A.3) are assumed to hold for $t = 0, 1$ instead for only $t = 0$, then C could be used instead of C_0 thereby increasing the power of the test. As a benchmark we also implement a parametric DWH test, where we first regress T on X and Z and then add the residuals from this fit as a covariate into the outcome equation for Y . The test for the unconfoundedness assumption is then a Wald test on the parameter for the included residual covariate (see, e.g. Wooldridge [26, Chap. 6], and Rivers and Vuong [27]). We use a robust covariance matrix [28].

3.1 Design

We consider a data generating process (DGP) which mimics a situation with a randomized assignment to a treatment (Z) with non-perfect compliance ($\delta_0 = 0$ below), where T denotes the actual treatment assignment, as well as more general situations where the effect of Z on T is allowed to be confounded by unobservables. For unit i , let

$$Z_i = I(\delta_0 U_{0i} + \varepsilon_{Zi} > 0),$$

$$T_i = I(X_i + \delta_1 U_{0i} + (0.5 + \delta_2 U_{1i})Z_i + U_{2i} + \varepsilon_{Ti} > 0),$$

and

$$Y_i = 1 + X_i + \theta_i T_i + \delta_3 U_{2i} + \varepsilon_{Yi}$$

or

$$Y_i = I(1 + X_i + \theta_i T_i + \delta_3 U_{2i} + \varepsilon_{Yi} > 0).$$

Table 1 Different designs considered with resulting instrumental property for Z and whether nonparametric identification of the (local) average causal effect holds

	DGP		Identification*
	$Y_i \in$	Parameter values	
D.1	$\mathbb{R}$	$\delta_0 = \delta_1 = 0, \delta_2 = 0$	Yes
D.2	$\mathbb{R}$	$\delta_0 = \delta_1 = 0, \delta_2 = 8$	No
D.3	$\mathbb{R}$	$\delta_0 = 1, \delta_1 = 0.2, \delta_2 = 0$	No
D.4	$\{0, 1\}$	$\delta_0 = 1, \delta_1 = 0, \delta_2 = 0$	Yes
D.5	$\{0, 1\}$	$\delta_0 = 1, \delta_1 = 0.2, \delta_2 = 0$	No

Notes: *Nonparametric identification of local average causal effects holds with non-confounded instruments ($\delta_0 \times \delta_1 = 0$) which fulfill a monotonicity assumption ($\delta_2 = 0$ Angrist et al. [4]; Abadie [5]; Frölich [6]).

We let $\varepsilon_{Yi}, \varepsilon_{Zi}, \varepsilon_{T_i(0)}, \varepsilon_{T_i(1)}, U_{0i}, U_{1i}$ and U_{2i} be independently distributed as $N(0, 0.25)$. Moreover, we also let $X_i \sim N(0, 2)$, and consider two cases for θ_i : $\theta_i = 1$ (homogeneous treatment effect) and $\theta_i = 1 + X_i$ (heterogeneous treatment effect). Parameters are varied in the study in order to study the empirical size and power of the test C_0 . Five designs, denoted D.1–D.5, are considered and described in Table 1. For the situation where we set $\delta_2 = 8$ (Design D.2), the instrumental variable Z is non-monotone, i.e. there exists individuals j for which $T_j(Z_j = 0) = 1$ and $T_j(Z_j = 1) = 0$ (called defiers), where $T_j(Z_j = k)$, $k \in \mathcal{T}$, are potential treatment values for individual j when switching Z_j to (everything else equal) k equal 0 or 1. The proportion of defiers when $\delta_2 = 8$ is 8.4%. Thus, for design D.2 the monotonicity assumption necessary for the nonparametric identification of the local average causal effect is violated [4–6]. Another assumption for identification made in the latter references is that $\delta_0 \times \delta_1 = 0$, and, hence, the instrument does not recover identification in designs D.3 and D.5.

The two tests mentioned above – C_0 and DWH – are evaluated in testing the null hypothesis $\delta_3 = 0$, and empirical size and power of the tests are obtained by letting $\delta_3 \in \{0, 0.1, 0.2, 0.3, 0.6, 0.9, 1.5, 2\}$. K -nearest neighbor matching estimators with $K = 1, 3, 5$ and 7 are used to compute C_0 , and we restrict X to have common support when conditioning on $Z = 1$ and $Z = 0$. We consider sample sizes $N = 500, 1,500$ and 3,000. In the continuous response cases, DWH should have correct size when $\theta_i = 1$ irrespective of whether the instrument is monotone or not, or whether the relation with T is confounded or not. DWH is also expected to have correct size [27] in the binary response case with homogeneous causal effect ($\theta_i = 1$). In contrast, DWH is expected to breakdown in all heterogeneous cases ($\theta_i = 1 + X_i$), since the response model is then misspecified. Up to our knowledge, no nonparametric test has previously been proposed in the literature for situations in Table 1 where an average causal effect is not nonparametrically identified. On the other hand, using C_0 is expected to give correct size and has power in all situations simulated.

3.2 Results

The results from the Monte Carlo simulations are displayed in Figures 3 and 4. The empirical sizes are also displayed in Table 2. The nonparametric test C_0 with $K = 5$ behaves well with all the DGPs considered, with empirical size close to 5% and power increasing with sample size. Results with other values for K can be obtained from the authors. Empirical sizes were comparable for all K values considered, while power increased with K : significantly so from $K = 1$ to $K = 3$ and only marginally from $K = 5$ to 7. Power is further increased when using C instead of C_0 (see Table 3 for design D.1; similar increase was obtained for the other designs) as expected since the former is based on stronger assumptions. On the other hand, the DWH test has too large empirical size in the heterogeneous cases ($\theta_i = 1 + X_i$). In the homogeneous treatment setup ($\theta_i = 1$) DWH behaves well with respect to its empirical size. This was expected as noted in the previous section, thereby yielding an interesting benchmark. In such homogeneous cases, the nonparametric test C_0

Figure 3 Empirical size ($\delta_3 = 0$) and power for the nonparametric test C_0 and the DW(H) test (based on robust covariance matrix) for Design D.1 (first row) and Design D.2 (second row), homogeneous causal effect (first column) and heterogeneous causal effect (second column). Designs are described in Table 1

Figure 4 Empirical size ($\delta_3 = 0$) and power for the nonparametric test C_0 and the DW(H) test [27] for Design D.3–D.5 (rows 1–3), homogeneous causal effect (first column) and heterogeneous causal effect (second column). Designs are described in Table 1

Table 2 Empirical sizes (nominal size is 5%) obtained with the nonparametric test C_0 ($K = 5$) and the DWH test for simulated DGPs with a continuous response

	$Y_i \in$	DGP				DWH			C_0		
		θ_i	δ_0	δ_1	δ_2	Sample size			Sample size		
						500	1,500	3,000	500	1,500	3,000
D.1	$\mathbb{R}$	1	0	0	0	5.46	4.55	4.94	5.47	5.14	5.24
	$\mathbb{R}$	$1 + X_i$	0	0	0	5.85	6.15	7.79	5.47	5.16	5.22
D.2	$\mathbb{R}$	1	0	0	8	5.42	5.24	5.21	5.49	5.72	5.65
	$\mathbb{R}$	$1 + X_i$	0	0	8	99.51	100	100	5.50	5.73	5.65
D.3	$\mathbb{R}$	1	1	0.2	0	5.11	4.72	4.87	5.67	5.26	5.27
	$\mathbb{R}$	$1 + X_i$	1	0.2	0	5.66	6.62	8.18	5.67	5.28	5.26
D.4	$\{0, 1\}$	1	1	0	0	3.98	4.81	5.20	5.35	4.83	5.09
	$\{0, 1\}$	$1 + X_i$	1	0	0	5.47	6.09	7.43	5.35	4.85	5.09
D.5	$\{0, 1\}$	1	1	0.2	0	4.09	4.65	4.98	5.47	5.02	5.18
	$\{0, 1\}$	$1 + X_i$	1	0.2	0	5.75	6.99	8.70	5.47	5.03	5.18

Note: 95% confidence intervals for the empirical sizes are $\pm 1.9\%$ (500 replicates), $\pm 1.1\%$ (1,500) and $\pm 0.8\%$ (3,000).

Table 3 Empirical size (nominal size is 5%) and power obtained with test statistics C_0 and C (both with $K = 5$) for simulated DGP D.1 with $\theta_i = 1 + X_i$; sample size 3,000

δ_3	0	0.1	0.2	0.3	0.6	0.9	1.5	2
C_0	5.22	6.79	9.83	14.78	35.72	55.23	75.72	81.90
C	5.25	7.61	14.00	23.51	58.59	81.84	91.30	97.55

has similar or better power than DWH, except for Designs D.1, where DWH is based on correctly specified models. For Design D.2 (non-monotone instrument), C_0 has markedly higher power than DWH.

In summary, the results obtained show that the nonparametric test (3) performs well in situations where DWH is consistent. By making fewer assumptions, (3) is also shown to work with non-monotone instruments and instruments whose effect on the treatment is confounded by unobservables, i.e. in situations where a local average causal effect is not identified.

4 Effect of JP

We consider a case study where the interest lies in estimating the effect of JP for unemployed on employment status. JP was offered within two separate labor market training (LMT) programs in Sweden during 1998. One program was run by the regular program provider in Sweden; the Swedish National Labor Market Board (AMV). The other program was offered by the Federation of Swedish Industries (Swit). To be eligible to the programs the unemployed individuals had to be at least 20 years of age and enrolled at the public employment service. There was no difference in benefits for the two groups of trainees. The fundamental idea with the Swit program was to increase the contacts between the unemployed individuals and employers by providing JP. In a survey conducted in June 2000 on 1,000 program participants from both programs, 69.5% of the Swit participants and 52% of the AMV participants stated that they obtained access to JP.³ Except for the idea to provide more contacts with employers the two programs were similar. Both programs

³ A detailed description of the survey can be found in Johansson and Martinson [29]. The survey contained a total of 19 questions. These concerned (i) the individual's background, (ii) the individual's labor market training and (iii) the individual's present labor market situation.

Table 4 The frequency distribution of the courses within the two programs

	AMV (<i>n</i> = 796)	Swit (<i>n</i> = 794)
Programmer	32	27
Computer technician	31	29
Application support	10	16
IT-pedagogue	2	6
IT-entrepreneur	1	3
Other	17	15
Missing	7	4

tested the individual's motivation and ability before recruitment by similar selection procedures (see Johansson [30], for a thorough description of the selection). The types of courses given within the Swit and the AMV programs are displayed in Table 4. The similarities of the two programs are apparent. Thus, despite differences in procurement between the two organizations (Swit and AMV), there do not seem to be any large differences between the types of LMT courses offered nor with the selection of participants. The fact that the programs distinguish themselves only with respect to JP availability prompts us that the effect of LMT program choice on labor market outcome should differ only through the effect of JP. This suggests that LMT program choice has the property (A.2) of an instrument for JP.

Based on the survey one can see in Table 5 that there is a statistical significant 18.1 percentage points difference in employment six months after leaving the program (the two programs have same average length) when comparing individuals having JP with those without. In the table we have some individual background variables: (i) education, (ii) work handicap (see disabled), (iii) gender (1 if man and 0 if women) and (iv) immigration status (1 if immigrant 0 else). Finally, since the propensity of receiving JP are higher in larger labor markets with also better labor market opportunities we need to control for region of residence in the estimation of an effect of JP. Sweden was divided into four regions: Stockholm, Skåne, Västra Götaland and the rest of the country. Stockholm, Skåne and Västra Götaland are the three regions with the largest population. Note that we have good reasons to assume that the two LMT programs only differ in their JP prospects, thus if the labor market opportunities affect the access to the LMT programs this does not invalidate them being used as an instrument for JP.

We can see some average differences between the two samples. Those with JP are (i) less disabled and (ii) less likely to live in Stockholm. The level of education also differs: they have on average more compulsory and upper secondary education but also less college education than those with no JP. Based

Table 5 Descriptive statistics for outcome employment and background characteristics and how they differ between JP and non-JP individuals

JP	Yes	No	Diff	t-test
Employment	64.9	46.8	18.1	6.82
Compuls. educ.	5.1	7.6	-2.5	-1.9
Upper sec. educ.	67.8	62.1	5.7	2.19
College	27.1	30.3	-3.2	-1.3
Disabled	7.5	11.5	-4.0	-2.5
Man	62.1	61.9	0.2	0.1
Immigrant	5.6	6.4	-0.9	-0.7
Stockholm	21.4	27.8	-6.5	-2.7
Skå ne	10.6	8.3	2.3	1.5
Västra Götaland	13.8	16.5	-2.6	-1.3
Rest of the country	54.2	47.3	2.7	2.53
Sample size	969	528

Figure 5 The distribution (percent) of the estimated probabilities (as function of the covariates) of (not) having JP (T , left panel) and of getting into the two alternative LMT programs (Z , right panel)

on these average differences, it is difficult to argue that those with JP have better labor market prospects than those without JP. The single factor suggesting the JP population has better labor opportunities without JP is that they are less likely disabled. In order to further study the selection into JP we used the covariates from the table and estimated a logistic regression model (a propensity score) including merely main effects. The results from this estimation (not displayed) are that individuals who are from Stockholm or Västra Götaland, and disabled, are less likely to receive JP. There is no statistical significant (5% level) differences in education between the two groups for instance. Figure 5, left panel, displays the propensity scores estimated. The latter gives evidence for the common support assumption in (A.1). In order to investigate the related assumption $0 < \Pr(Z = 1|\mathbf{X}) < 1$ included in (A.2), we also fit the probability of getting into Swit versus AMV with a logistic regression including main effects, and Figure 5, right panel, also provides evidence for the latter assumption.

Because there are 969 JP (treated) individuals for only 528 non-treated individuals an estimate of the average causal effect of JP on the treated (ACT), θ^t , will typically suffer from severe bias due to difficulties in finding matches to the treated. Thus, we estimate instead the average causal effect of JP on the non-treated (ACNT), θ^{nt} . A reasonable assumption is that individuals with higher than average return from JP are the ones who select themselves into JP. This means that ACNT yields a lower bound for ACT, $\theta^t \geq \theta^{nt}$.

Assumption (A.1) need only to be fulfilled for $t = 1$ in order for us to estimate ACNT, i.e. $Y(1) \perp\!\!\!\perp T|\mathbf{X}$, where the covariates are displayed in Table 5. A $K = 5$ nearest neighbor matching estimator using the minimum Mahalanobis distance between the covariates of Table 5 is used to estimate the parameter θ^{nt} , yielding $\hat{\theta}^{nt} = 12\%$ points, with standard error [23, Theorems 6 and 7] estimated to 5% points. Hence, there is a significant effect from JP.

4.1 Testing the unconfoundedness assumption

We test for the null hypothesis H_0^b using C_1 in (3). Nonparametric estimation is performed with $K = 5$ nearest neighbor matching on the covariates displayed in Table 5 using the minimum Mahalanobis distance, also for computing the standard deviation s_1 ; see Abadie and Imbens [23]. The resulting value for test statistic is 1.31. Hence, we cannot reject the unconfoundedness assumption (p -value of 0.18). We also perform a DWH test by estimating a linear probability model with the discrete covariates displayed in Table 5, yielding a p -value of 0.09. Thus, given the maintained assumption (A.2), none of the test can reject the null hypothesis, at the 5% level, that the effect of JP on employment is not confounded, although the DWH test by making stronger assumptions has a p -value under 10%.

5 Conclusions

Identification of the causal effect of a treatment on an outcome in observational studies is typically based either on the unconfoundedness assumption or on the availability of an instrument (e.g. Angrist et al. [4]). In this paper, by introducing general instrumental assumptions we are able to propose an easy to use nonparametric test for the unconfoundedness assumption in situations where the same assumptions do not allow for the nonparametric identification of a causal effect. We illustrate the framework introduced with a study of the effect of JP for unemployed on employment, where we argue that an instrument fulfilling our conditions is available through the existence of two LMT programs with different degree of accessibility to JP.

In many applications, nonparametric identification of causal effects using instruments is non-trivial, e.g. when a non-testable monotonicity property for the instrument must hold [4–6] and/or when a large set of control variables is needed for the instrument to be valid. Using our weaker instrumental conditions, one may test for the unconfoundedness assumption. If the latter is not rejected, this gives some ground to the analyst to proceed using an identification strategy based on the unconfoundedness assumption. We have operationalized the theoretical results with a test statistic based on K -nearest neighbor matching estimators. Other nonparametric regression estimators may be used instead, such as, e.g. local polynomial regression and splines. Finally, it is worth noting here that for durations outcomes with censored data, the test proposed herein may be implemented by making use of the matching estimators for censored duration responses presented in Fredriksson and Johansson [31] and de Luna and Johansson [32].

Acknowledgments: This paper has benefited from useful comments from Martin Huber, Ingeborg Waernbaum, an editor, an anonymous referee and seminar participants at John Hopkins, Maryland University and the third Joint IZA/IFAU Conference on Labor Market Policy Evaluation. De Luna acknowledges the financial support of the Swedish Research Council through the Swedish Initiative for Research on Microdata in the Social and Medical Sciences (SIMSAM), the Ageing and Living Condition Program and grant 70246501. Johansson acknowledges the financial support of the Swedish Council for Working Life and Social Research (grant 2004–2005).

References

1. de Luna X, Johansson P. Exogeneity in structural equation models. *J Econometrics* 2006;132:527–43.
2. Imbens GW, Wooldridge JM. Recent developments in the econometrics of program evaluation. *J Econ Lit* 2009;47:5–86.
3. Pearl J. *Causality*, 2nd ed. Cambridge: Cambridge University Press, 2009.
4. Angrist D, Imbens G, Rubin D. Identification of treatment effects using instrumental variables. *J Am Stat Assoc* 1996;91:444–55.
5. Abadie A. Semiparametric instrumental variable estimation of treatment response models. *J Econometrics* 2003;113:231–63.
6. Frölich M. Nonparametric iv estimation of local average treatment effects with covariates. *J Econometrics* 2007;139:35–75.
7. Hahn J, Todd P, van der Klaaw W, Todd W, Van der Klaauw P. Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica* 2001;69:201–9.
8. Battistin E, Retore E. Ineligibles and eligible non-participants as a double comparison group in regression discontinuity designs. *J Econometrics* 2008;142:715–30.
9. Dias M, Ichimura H, van den Berg G. The matching method for treatment evaluation with selective participation and ineligibles. IFAU Working Papers, 2008:6, Institute for Labour Market Policy Evaluation, Uppsala, 2008.
10. Lee M-J. *Micro-econometrics for policy, program and treatment effects*. Oxford: Oxford University Press, 2005.
11. Donald SG, Hsuz Y-C, Lieli RP. Testing the unconfoundedness assumption via inverse probability weighted estimators of (I) att. Working Paper, 2011.
12. Imbens GW, Angrist JD. Identification and estimation of local average treatment effects. *Econometrica* 1994;62:467–75.
13. Angrist J, Fernandez-Val I. *ExtrapLATE-ing: external validity and overidentification in the late framework*. NBER Working Paper, 16566, National Bureau of Economic Research, Cambridge, MA, 2010.

14. Guo Z, Cheng J, Lorch S, Small D. Using an instrumental variable to test for unmeasured confounding. Working Papers, 2013.
15. Rosenbaum PR. The role of a second control group in an observational study (with discussion). *Stat Sci* 1987;2:292–316.
16. Neyman J. Sur les applications de la théorie des probabilités aux expériences agricoles: essai des principes. *Roczniki Nauk Rolniczych X* 1923:1–51. In Polish, English translation by D. Dabrowska and T. Speed in *Stat Sci* 1990;5:465–72.
17. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol* 1974;66:688–701.
18. Imbens GW. Nonparametric estimation of average treatment effects under exogeneity: a review. *Rev Econ Stat* 2004;86:4–29.
19. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983;70:41–55.
20. de Luna X, Waernbaum I, Richardson T. Covariate selection for the non-parametric estimation of an average treatment effect. *Biometrika* 2011;98:861–75.
21. Dawid AP. Conditional independence in statistical theory. *J R Stat Soc Ser B* 1979;41:1–31.
22. Lauritzen S. Graphical models. Oxford: Oxford University Press, 1996.
23. Abadie A, Imbens GW. Large sample properties of matching estimators for average treatment effects. *Econometrica* 2006;74:235–67.
24. de Luna X, Johansson P, Sjöstedt-de Luna S. Bootstrap inference for k -nearest neighbour matching estimators. IZA Discussion Papers 5361, Institute for the Study of Labor, Bonn, 2010.
25. Su L, White H. A consistent characteristic function-based test for conditional independence. *J Econometrics* 2007;141:807–34.
26. Wooldridge J. Econometric analysis of cross section and panel data. Cambridge: MIT Press, 2002.
27. Rivers D, Vuong H. Limited information estimators and exogeneity tests for simultaneous probit models. *J Econometrics* 1988;39:347–66.
28. White H. Maximum likelihood estimation of misspecified models. *Econometrica* 1982;50:1–25.
29. Johansson P, Martinson S. Det nationella it-programmet – en slutrapport om swit. Forskningsrapporter, 2000:8, Institute for Labour Market Policy Evaluation, Uppsala, 2000.
30. Johansson P. The importance of employer contacts: evidence based on selection on observables and internal replication. *Labour Econ* 2008;15:350–69.
31. Fredriksson P, Johansson P. Dynamic treatment assignment – the consequences for evaluations using observational data. *J Bus Econ Stat* 2008;26:435–45.
32. de Luna X, Johansson P. Non-parametric inference for the effect of a treatment on survival times with application in the health and social sciences. *J Stat Plann Inference* 2010;140:2122–37.