Testing for Nonparametric Identification of Causal Effects in the Presence of a Quasi-Instrument

Xavier de Luna
Per Johansson

Testing for Nonparametric Identification of Causal Effects in the Presence of a Quasi-Instrument

Xavier de Luna
Umeå University

Per Johansson
Uppsala University,
IFAU and IZA

ABSTRACT

Testing for Nonparametric Identification of
Causal Effects in the Presence of a Quasi-Instrument[*]

The identification of average causal effects of a treatment in observational studies is typically based either on the unconfoundedness assumption or on the availability of an instrument. When available, instruments may also be used to test for the unconfoundedness assumption (exogeneity of the treatment). In this paper, we define variables which we call quasi-instruments because they allow us to test for the unconfoundedness assumption although they do not necessarily yield nonparametric identification of the average causal effect. A quasi-instrument is defined as an instrument except for that its relation to the treatment is allowed to be confounded by unobservables, thereby resulting in a wider range of potential applications. We propose a test for the unconfoundedness assumption based on a quasi-instrument, 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.

Corresponding author:

Per Johansson
IFAU
Box 513
SE-75120 Uppsala
Sweden
E-mail: per.johansson@ifau.uu.se

---

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, 2006, and the references therein) 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 which 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 and hence also to test the unconfoundedness assumption. Such a test is typically performed by comparing the estimates of the causal effects obtained both under the unconfoundedness assumption, and by using the instrument (classical Durbin-Wu-Hausman test). Nonparametric identification is also possible with the help of instruments and, for instance, Angrist et al. (1996) develop a theory for the nonparametric identification and estimation of local average causal effects. Based on these results, Donald et al. (2011) propose a test of the unconfoundedness assumption. Abadie (2003) and Frölich (2007) extended Angrist et al. (1996) results to the situation where the observed covariates are related to the instrument. Non-parametric identification can also be obtained with the related concept of (fuzzy) regression discontinuity designs; see Hahn et al. (2001), Battistin and Retore (2008), Dias et al. (2008) and Lee (2005, Sec. 5.5.3).

In this paper, we define quasi-instruments, i.e. variables with the properties of an instrument except that their effect on the treatment T is allowed to be confounded by unobserved variables. Therefore, in general, quasi-instruments will not yield identification of a causal effect when the the unconfoundedness assumption does not hold. An exception to this is the case of parametric linear systems where quasi-instruments yield identification (Pearl, 2009, p. 248). On the other hand, we show that quasi-instruments (which are based on weaker assumptions than instruments) can be used to obtain test statistics for the unconfoundedness hypothesis. 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 (1987), de Luna and Johansson (2006) and Dias et al. (2008). Rosenbaum (1987) was probably first to formalize the idea that two control groups provide information on the unconfoundedness assumption and described actual observational studies where different controls groups where available. One of our contributions in this context is the introduction of mild assumptions (defining quasi-instruments) under which

an available control group may be split into two to test the unconfoundedness assumption non-parametrically, although the test statistic we eventually propose does not actually require the split to be done. The nonparametric identification of local average causal effects with instruments and the related test proposed by Donald et al. (2011) require not only the instrument to be uncounfounded, as mentioned above, but also that it affects treatment in a monotone fashion as defined, e.g., in Imbens and Angrist (1994), Angrist et al. (1996) and Angrist and Fernandez-Val (2010). We, thus, contribute to this literature because the test we introduce is based on an instrument for which the untestable monotonicity assumption does not need to hold.

In Section 4 we present a motivating example where Swedish register data is used to study the causal effect of job practice on employment. We have access to a rich set of background characteristics on unemployed individuals, although the question remains whether the effect of job practice on employment is confounded by unobserved heterogeneity. In this study, unemployed have access to job practice through their participation into a labor market program. During 1998 there were two such labor market programs available in Sweden offering job practice with different probabilities. Because we have evidence that the two programs differ mainly only with respect to their propensity to offer job practice, the participation into the two programs may be assumed to affect employment differently only through job practice. We, thus, consider program participation to be a quasi-instrument. In contrast with usual instrumental assumptions this allows unobserved heterogeneity to confound the effect of program participation on the probability to be offered job practice. Finally, we apply the proposed test using the quasi-instrument to check whether the effect of job practice on employment is confounded by unobservables.

Before treating this motivating example in more details in Section 4, Section 2 presents the model, defines quasi-instruments, and develops the theoretical results which allow us to then introduce a test of the unconfoundedness assumption. Section 3 presents a simulation study of finite sample performances of the proposed test. In particular, one of the design 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 (Neyman, 1923; Rubin, 1974) 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 Z two vectors of such variables, the latter of dimension one.

For $t \in \mathcal{T}$, we consider $(\mathbf{X}, Z, T, Y(t))$ 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 $E(Y(1) - Y(0))$ and the average causal effect on the treated $E(Y(1) - Y(0) \mid T = 1)$.

In observational studies, where the treatment assignment T is not randomized by definition, an identifying assumption (e.g., Rosenbaum and Rubin, 1983; Imbens, 2004) for the average causal effect is the following.

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

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

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. However, this identifiably assumption is untestable without further assumptions and/or information.

2.2 Quasi-instrument, 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 $t \in \mathcal{T}$,

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

Assumption (A.2) prohibits, e.g., a direct effect from Z to Y, i.e. an effect not going through T, or unobserved variables affecting both Z and Y. We also need the following regularity condition.

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

Note that assumption (A.3) is not equivalent to assuming $(Z, T) \perp\!\!\!\perp Y(t)|\mathbf{X}$, and the subtle distinction is necessary in Proposition 2 below. We have also that because $(Z, T) \perp\!\!\!\perp Y(t)|\mathbf{X}$ implies (A.1) and (A.2), (A.3) says that (A.1-2) holds if and only if $(Z, T) \perp\!\!\!\perp Y(t)|\mathbf{X}$ holds. That joint independence, $(Z, T) \perp\!\!\!\perp Y(t)|\mathbf{X}$, is implied by the marginal independencies, (A.1) and (A.2), does not hold necessarily. On the other hand, a sufficient condition for assumption (A.3) to hold is, for instance, that $T \perp\!\!\!\perp Y(t)|Z, \mathbf{X}$ by the contraction property (Pearl, 2009, p. 11, and Dawid, 1979). In other words, (A.3) holds if the treatment assignment is independent of each potential outcomes when conditioning not only on $\mathbf{X}$ but also on Z. Example 1 provides a situation where (A.3) does not hold.

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:*

$$
\begin{aligned}
Z^* &= \psi_0 + \psi_1 U + \psi_2 V + \varepsilon_Z, \\
T^* &= \nu_0 + \nu_1 V + \varepsilon_T, \\
Y(0) &= \xi_0 + \xi_1 Z^* + \xi_2 U + \varepsilon_Y.
\end{aligned}
\tag{1}
$$

where $U, V, \varepsilon_Z, \varepsilon_T, \varepsilon_Y$ 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
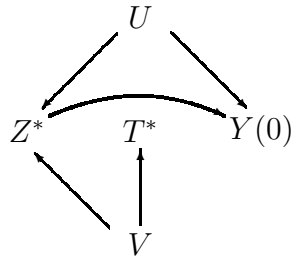

Figure 1: Graph illustrating model (1) in Example 1.

a graphical representation of the model, where $\varepsilon_Z, \varepsilon_T, \varepsilon_Y$ are omitted. We can write the conditional expectations

$$E(Y(0) \mid Z^*, U) = \alpha Z^* + \beta U,$$
$$E(U \mid Z^*) = \gamma Z^*,$$

where the parameters in the latter are function of the parameters of the joint normal distribution.

In Example 1, (A.1) holds irrespective of the parametrization. In contrast, (A.2) will typically be violated, unless we assume that $\alpha = -\beta\gamma$, in which case $Z^* \perp\!\!\!\perp Y(0)$ by joint normality, and hence $Z \perp\!\!\!\perp Y(0)$. The constrained parametrization $\alpha = -\beta\gamma$, yields thus an example where (A.3) is violated since (A.1) and (A.2) hold while $(Z, T) \perp\!\!\!\perp Y(0)$ does not necessarily hold. This type of example are called unstable (Pearl, 2000, Sec. 2.4) in the sense that (A.2) will ceased to hold as soon as the parameter values do not fulfill the constraint $\alpha = -\beta\gamma$.[1] Examples where (A.3) does not hold are typically unstable in this sense. Note that stability of the conditional independence statements is also necessary for model (1) to be identifiable. Stability assumptions are typically implicit in structural models.

Let us now state a first result.

Proposition 1 *Assume (A.1-3), then*

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

Proof. Under (A.1-3) we have that $(Z, T) \perp\!\!\!\perp Y(t) | \mathbf{X}$, $t \in \mathcal{T}$, hold. Then, the result of the proposition is a direct consequence of the weak union property (Pearl, 2009, p. 11, and Dawid, 1979) ∎

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.

---

[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 (1996) for a general account on graphical models and de Luna et al. (2011) for their use together with potential outcomes. Briefly, a distribution over the variables in a directed acyclic graph is stable for the graph if there is a one-to-one mapping between the conditional independence statements holding for the distribution and the conditional independence statements that can be read from the graph. Conditional independence statements are read from a graph as follows: $T \perp\!\!\!\perp Y(0) | V$ holds for the graph if all the connected paths between T and $Y(t)$ go through V, and V is not a collider on one of the paths, where V is a collider if $A \longrightarrow V \longleftarrow B$ with A and B the two neighbours of V on the path.

Thus, evidence against (2) can be interpreted as evidence against the unconfoundedness assumption (A.1) if (A.2-3) are known to hold from subject-matter considerations. One application is a random experiment (where Z is a random assignment) with restricted compliance and detailed information in the data on the individuals participating. In such a case the instrument could be used to test whether the information on participants contains all potential confounding heterogeneity. If this happen to be the case one may obtain identification without randomization in full scaled observational studies where the same baseline information is available. Another example of application is treated in detail in Section 4.

For a test based on (2) to have power against (A.1) we further need to assume:

(A.4) Z and T are dependent conditional on $\mathbf{X}$.

This assumption is typically made for instrumental variables to be useful for identification. The following result shows that (A.4) is a necessary condition for a test based on (2) to have power against the alternative hypothesis to (A.1): T and $Y(t)$ are dependent conditional on $\mathbf{X}$.

Proposition 2 *Assume (A.2-3). Then,*

$$\{(2) \implies (A.1)\} \implies (A.4).$$

Proof.   Let us show the contrapositive and assume that (A.4) does not hold, i.e. Z and T are independent given $\mathbf{X}$. Then, there exists situations where (2) holds and yet (A.1) does not hold; see Figure 2, where such a case is described with a directed acyclic graph. Hence, we have $\neg\{(2) \implies (A.1)\}$, where $\neg A$ is the negation of statement A, thereby showing the proposition. ∎
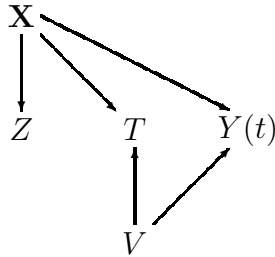

Figure 2: This figure displays a directed acyclic graph which together with a stable distribution for the variables (see Footnote 1) give an example where (2) holds and yet (A.1) does not hold. This is used in the proof of Proposition 2.
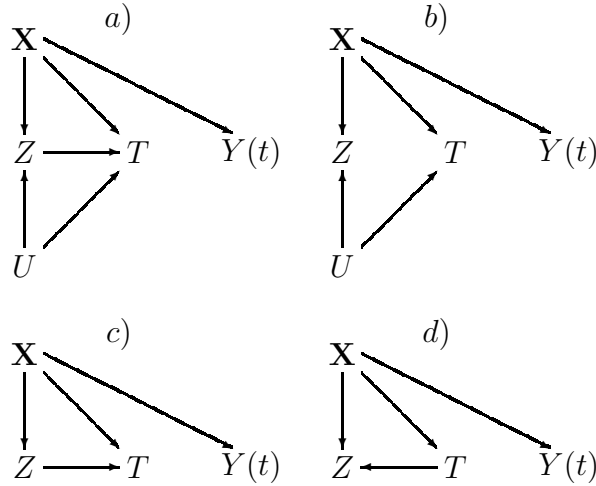
7

Figure 3: All four directed acyclic graphs together with a respective stable joint distribution for the variables included are examples where (A.4) holds. However, only cases a), b) and c) are such that the test based on (2) has power, i.e. if (A.1) does not hold, e.g. through the introduction of a variable V with arrows pointing towards T and $Y(t)$, then $Y(t) \perp\!\!\!\perp Z \mid T, \mathbf{X}$ would not hold either.

Sufficient conditions (expressed with directed acyclic graphs; see Footnote 1) for a test based on (2) to have power against (A.1) are given in Figure 2, panels a), b) and c), while panel d) shows a case where the test does not have power.

Identification of the effect of T on Y is guaranteed under (A.2) and (A.3-4) with linear models, see, e.g., Pearl (2009, p. 248). This, however, is not true in general (even for nonparametric identification of local average causal effects; Angrist et al., 1996; Frölich, 2007) and we therefore call a variable Z not affected by T and for which (A.2-4) hold a *quasi*-instrument.

2.3 Method

For the sake of simplicity, we consider the situation where the parameter of interest is the average causal effect on the treated, $\theta = E(Y(1) - Y(0)|T = 1)$. The results are symmetric when the interest lies on the average causal effect on the non treated. Here, assumptions (A.1-3) need to hold only for $t = 0$. Different strategies may be adopted to test the null hypothesis defined by the conditional independence statement of Proposition 1 with $t = 0$, i.e.

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

One strategy could be to use the concept of two independent control groups (Rosenbaum, 1987). Under H_0 we can use Z to obtain two independent control groups (one defined by $Z = 1$ and one by $Z = 0$) for estimating θ, yielding $\widehat{\theta}^{z=0}$ and $\hat{\theta}^{z=1}$. Under H_0 the difference $\widehat{\theta}^{z=0} - \hat{\theta}^{z=1}$ should have mean zero and this is the base for a test statistic. However, since we need to compute two non-parametric 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.[2]

In this paper we propose a testing strategy based on the fact that under H_0 we have $\delta(\mathbf{X}) = 0$, for all $\mathbf{X}$, where

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

Consider a non-parametric estimator for $\delta = E(\delta(\mathbf{X}))$,

$$\hat{\delta} = \sum_{i:Z=1} [Y_i(0) - \hat{Y}_i(0)] + \sum_{i:Z=0} [Y_i(0) - \tilde{Y}_i(0)],$$

where $\hat{Y}_i(0)$ is a non-parametric estimator of $E(Y_i(0) \mid T_i = 0, \mathbf{X}_i, Z_i = 0)$ and $\tilde{Y}_i(0)$ is a non-parametric estimator of $E(Y_i(0) \mid T_i = 0, \mathbf{X}_i, Z_i = 1)$. The two latter estimates may be obtained by nearest neighbour matching, or any other smoothing technique. Since $\delta = 0$ under H_0, the test statistic

$$C = \frac{\hat{\delta}}{s} \tag{3}$$

will, under the necessary regularity conditions, be normally distributed with mean zero and variance one, where s is the standard error of $\hat{\delta}$. For instance, if nearest neighbour matching estimators are used, then s was given in Abadie and Imbens (2006, Theorems 6 and 7). A subsampling estimator is also available in this case in de Luna et al. (2010).

3 Monte Carlo study

We use a Monte Carlo study to investigate the finite sample properties (empirical size and power) of the test (3) where nearest neighbour matching (on common support) is used as non-parametric estimator of $\hat{Y}_i(0)$ and $\tilde{Y}_i(0)$. As a benchmark we also implement a parametric Durbin-Wu-Hausman (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

---

[2]This has been confirmed in simulation experiments not presented here.

is then a Wald test on the parameter for the included residual covariate (see, e.g., Wooldridge, 2002, Chap. 6, and Rivers and Vuong, 1988). We use a robust covariance matrix (White, 1982).

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, where T denotes the actual treatment assignment, as well as more general situations where the effect of Z on Y 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

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

We let ε_{Yi}, ε_{Zi}, $\varepsilon_{T_i(0)}$, $\varepsilon_{T_i(1)}$, 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). Moreover, we also let $X_i \sim N(0, 2)$, and consider two cases for θ_i: $\theta_i = 1$ and $\theta_i = 1 + X_i$. Parameters are varied in the study in order to study the empirical size and power of test (3) in different situations. Five designs, denoted D.1 to D.5, are considered and described in Table 1. Thus, 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 $T_j(Z_j = k)$, $k = 0, 1$, 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% instead of zero for the monotone case.

The two tests mentioned above $-(3)$ and DWH$-$ are evaluated in testing the null hypothesis $\delta_3 = 0$, and empirical size and power of the tests are obtained by letting $\delta_3 \in \{0, 0.1, 0.2, 0.3, 0.6, 0.9, 1.5, 2\}$. We consider sample sizes $N = 500, 1500$ and 3000. 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 (quasi-instrument) or not. DWH test is also expected to have correct size (Rivers and Vuong, 1988) in the binary response case with homogenous causal effect $(\theta_i = 1)$. In contrast, DWH is expected to

Table 1: Different designs considered with resulting instrumental property for Z and whether no n-parametric identification of the (local) average causal effect hold.

	$Y_i \in$	DGP parameter values	Z_i	identification*
D.1	$\mathbb{R}$	$\delta_0 = \delta_1 = 0, \delta_2 = 0$	monotone instrument	yes
D.2	$\mathbb{R}$	$\delta_0 = \delta_1 = 0, \delta_2 = 8$	non-monotone instrument	no
D.3	$\mathbb{R}$	$\delta_0 = 1, \delta_1 = 0.2, \delta_2 = 0$	quasi-instrument	no
D.4	$\{0, 1\}$	$\delta_0 = 1, \delta_1 = 0, \delta_2 = 0$	monotone instrument	yes
D.5	$\{0, 1\}$	$\delta_0 = 1, \delta_1 = 0.2, \delta_2 = 0$	quasi-instrument	no

*Non-parametric identification of local average causal effects holds with non-confounded instruments ($\delta_0 \times \delta_1 = 0$) which fulfill a monotonicity assumption (Angrist et al., 1996; Frölich, 2007).

breakdown in all heterogeneous cases ($\theta_i = 1 + X_i$), since the response model is then misspecified. Up to our knowledge, no non-parametric test has previously been proposed in the literature for situations in Table 1 where an average causal effect is not non-parametrically identified. On the other hand, using test (3) is exepected to give correct size and have power in all situations simulated.

3.2 Results

The results from the Monte Carlo simulations are displayed in Figures 4 and 5. The empirical sizes are also displayed in Table 2. The non-parametric test (3) behaves well with all the data generating processes considered, with empirical size close to 5% and power increasing with sample size. On the other hand, the DWH test is not consistent (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 non-parametric test (3) has similar power than DWH, except for Designs D.1 and D.3 where DWH is based on correctly specified models yielding better power, and Design D.2 (non-monotone instrument) where DWH has markedly lower power.

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

11

Table 2: Empirical sizes (nominal size is 5%) obtained with the nonparametric test (3), denoted C, and the DWH test for simulated DGPs with a continuous response.

		DGP				DWH			C		
						Sample size			Sample size		
	$Y_i \in$	θ_i	δ_0	δ_1	δ_2	500	1500	3000	500	1500	3000
D.1	$\mathbb{R}$	1	0	0	0	5.46	4.55	4.94	5.39	5.10	4.93
	$\mathbb{R}$	$1 + X_i$	0	0	0	5.85	6.15	7.79	5.06	5.37	5.22
D.2	$\mathbb{R}$	1	0	0	8	5.42	5.24	5.21	5.25	5.14	5.24
	$\mathbb{R}$	$1 + X_i$	0	0	8	99.51	1.00	1.00	5.24	5.03	5.36
D.3	$\mathbb{R}$	1	1	0.2	0	5.11	4.72	4.87	5.57	5.08	5.35
	$\mathbb{R}$	$1 + X_i$	1	0.2	0	5.66	6.62	8.18	5.43	5.13	5.25
D.4	$\{0,1\}$	1	1	0	0	3.98	4.81	5.20	5.42	4.76	5.05
	$\{0,1\}$	$1 + X_i$	1	0	0	5.47	6.09	7.43	5.48	5.24	4.88
D.5	$\{0,1\}$	1	1	0.2	0	4.09	4.65	4.98	5.55	4.79	5.05
	$\{0,1\}$	$1 + X_i$	1	0.2	0	5.75	6.99	8.70	4.89	4.92	5.14

Note: 95% confidence intervals for the empirical sizes are: $\pm 1.9\%$ (500 replicates), $\pm 1.1\%$ (1500) and $\pm 0.8\%$ (3000).

4 Effect of job practice

We consider a case study where the interest lies in estimating the effect of job practice for unemployed on employment status. Job practice (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. From a survey conducted in June 2000 on $1,000$ program participants from both programs it could be seen that 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 tested the individual's motivation and ability before recruitment by similar selection procedures (see Johansson, 2008, for a thorough description of the selection). The types of courses given within the

---

[3]A detailed description of the survey can be found in Johansson and Martinson (2000). 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.
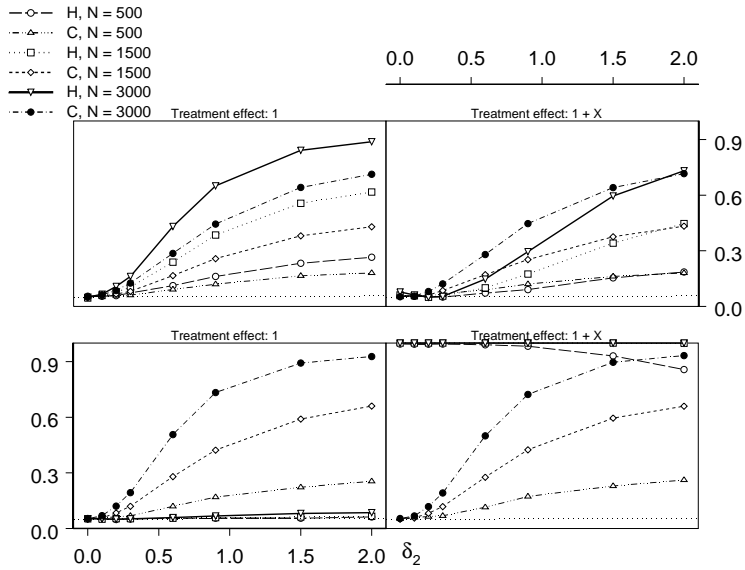
12

Figure 4: Empirical size ($\delta_3 = 0$) and power for nonparametric test (3), denoted C, 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.

Swit and the AMV programs are displayed in Table 3. The similarities of the two programs are apparent. Thus, despite the differences in procurement between the two organizations (Swit and AMV), there do not seem to be any large differences between the types of labor market training courses offered nor with the selection of participants. The fact that the programs distinguish themselves only with respect to job practice availability prompts us that the effect of LMT program choice on labor market outcome should differ only through the effect of job practice. This suggests that LMT program choice has the properties of a quasi-instrument as defined above.

Based on the survey one can see in Table 4 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 job practice 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, we have information on the individuals region of residence. Sweden was divided into four regions: Stockholm, Skåne, Västra Götaland and the rest of the
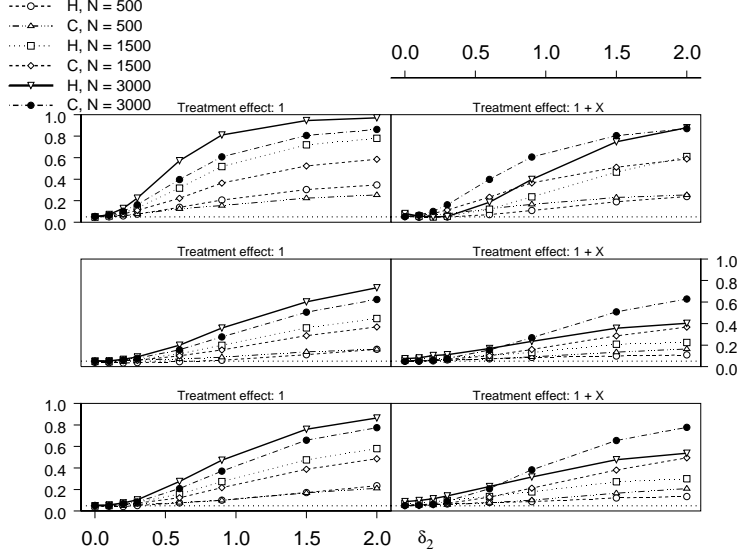
Figure 5: Empirical size ($\delta_3 = 0$) and power for the nonparametric test (3), denoted C, and the DW(H) test (Rivers and Vuong, 1988) for Design D.3-D.5 (row 1-3), homogeneous causal effect (first column) and heterogeneous causal effect (second column). Designs are described in Table 1.

country. Stockholm, Skåne and Västra Götaland are the tree regions with the largest population and with, in general, the best labor market opportunities.

We can see some average differences between the two samples. Those with job practice 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 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). 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 6, left hand panel, displays the propensity scores estimated. The latter gives evidence for the common support assumption in (A.1). In order to test the related assumption $\Pr(Z = t \mid \mathbf{X}) > 0$ included in (A.2), we also fit the

14

Table 3: The frequency distribution of the courses within the two programs.

	AMV ($n = 796$)	Swit ($n = 794$)
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

probability of getting into Swit versus AMV with a logistic regression, and Figure 6, right hand 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) 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), i.e.,

$$\tau = E(y(1) - y(0) \mid T = 0).$$

Because we expect positive selection into JP, ACNT yields a lower bound for ACT, $E(Y(1) - Y(0) \mid T = 1) \geq \tau$.

Assumption (A.1) need only to be fulfilled for $t = 1$ in order for us to estimate ACNT, in particular the unconfoundedness assumption needed here is

$$Y(1) \perp\!\!\!\perp T|\mathbf{X}, \tag{4}$$

where the covariates are displayed in Table 4.

A one-to-one matching estimator using the minimum Mahalanobis distance between the covariates of Table 4 is used to estimate the parameter τ, yielding $\hat{\tau} = 17\%$, with standard error (Abadie and Imbens, 2006, Theorems 6 and 7) estimated to 3%. Hence, there is quite a strong effect from JP.

4.1 Testing the unconfoundedness assumption

Here we test for the null hypothesis (4). Under this H_0 hypothesis we have that for all $\mathbf{X}$

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

Table 4: Descriptive statistics for outcome employment and background characteristics, and how they differ between JP and non JP individuals.

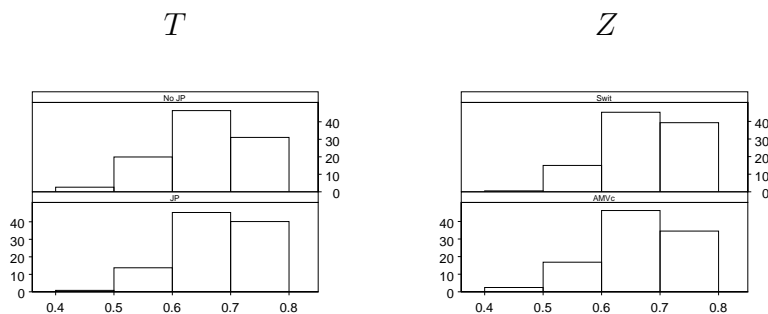| Job Practice	Yes	No	Diff	t-test
Employment	64.9	46.8	18.1	6.82
Compuls. educ.	5.1	7.6	-2.5	-1.9
Upper sec. educ.	67.8	62.1	5.7	2.19
College	27.1	30.3	-3.2	-1.3
Disabled	7.5	11.5	-4.0	-2.5
Man	62.1	61.9	0.2	0.1
Immigrant	5.6	6.4	-0.9	-0.7
Stockholm	21.4	27.8	-6.5	-2.7
Skåne	10.6	8.3	2.3	1.5
Västra Götaland	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 6: The distribution (percent) of the estimated probabilities (as function of the covariates) of (not) having JP (T, left hand panel) and of getting into the two alternative LMT programs (Z, right hand panel) .

16

The corresponding test statistic to (3) is then

$$\frac{\sum_{i:Z=1}[Y_i(1) - \hat{Y}_i(1)] + \sum_{i:Z=0}[Y_i(1) - \tilde{Y}_i(1)]}{s},$$

where $\hat{Y}_i(1)$ is a non-parametric estimator of $E(Y_i(1) \mid T_i = 0, \mathbf{X}_i, Z_i = 0)$ and $\tilde{Y}_i(1)$ is a non-parametric estimator of $E(Y_i(1) \mid T_i = 0, \mathbf{X}_i, Z_i = 1)$. Non-parametric estimation is obtained with one-to-one matching on the covariates displayed in Table 4 using the minimum Mahalanobis distance, also for computing the standard deviation s; see Abadie and Imbens (2006). The resulting value for test statistic (3) is 1.05. Hence, we cannot reject the unconfoundedness assumption (p-value of 0.29). We also perform a DWH test by estimating a linear probability model with the discrete covariates displayed in Table 4, yielding a p-value of 0.09. Thus, none of the test can reject the null hypothesis that the effect of job practice on employment is not confounded at the 5% level, although the DWH test by making stronger assumptions has a p-value under 10%.

5   Conclusion

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., 1996). In this paper, by introducing the new concept of quasi-instrument we are able to propose an easy to use non-parametric test for the unconfoundedness assumption. A quasi-instrument is a variable with the properties of an instrument except that their association with the treatment is allowed to be confounded by unobservables. Quasi-instruments fulfill weaker conditions than classical instruments, thereby the wider generality of the test introduced. We illustrate the framework introduced with a study of the effect of job practice for unemployed on employment, where we argue that a quasi-instrument is available through the existence of two labor market training programs with different degree of accessibility to job practice.

In many applications, non-parametric estimation of causal effects using instruments is non-trivial, e.g., when a non-testable monotonicity property for the instrument must hold (Angrist et al., 1996, Frölich,2007), and/or when a large set of control variables is needed for the instrument to be valid. Using weaker assumptions (quasi-instrument) one may test for the unconfoundedness assumptions allowing us to proceed using the unconfoundedness assumption when it cannot be rejected. 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 (2008) and de Luna and Johansson (2010).

References

Abadie, A. (2003). Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics 113*, 231–263.

Abadie, A. and G. W. Imbens (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica 74*, 235–267.

Angrist, D., G. Imbens, and D. Rubin (1996). Identification of treatment effects using instrumental variables. *Journal of the American Statistical Association 91*, 444–455.

Angrist, J. and I. Fernandez-Val (2010). Extrapolate-ing: External validity and overidentification in the late framework. *NBER Working Paper*, 16566, National Bureau of Economic Research, Cambrdige, USA.

Battistin, E. and E. Retore (2008). Ineligibles and eligible non-participants as a double comparison group in regression discontinuity designs. *Journal of Econometrics 142*, 715–730.

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society: Series B 41*, 1–31.

de Luna, X. and P. Johansson (2006). Exogeneity in structural equation models. *Journal of Econometrics 132*, 527–543.

de Luna, X. and P. Johansson (2010). Non-parametric inference for the effect of a treatment on survival times with application in the health and social sciences. *Journal of Statistical Planning and Inference 140*, 2122–2137.

de Luna, X., P. Johansson, and S. S. de Luna (2010). Bootstrap inference for k-nearest neighbour matching estimators. *IZA Discussion Papers* 5361, Institute for the Study of Labor, Bonn.

de Luna, X., I. Waernbaum, and T. Richardson (2011). Covariate selection for the non-parametric estimation of an average treatment effect. *Biometrika 98*, 861–875.

Dias, M., H. Ichimura, and G. van den Berg (2008). The matching method for treatment evaluation with selective participation and ineligibles. *IFAU Working Papers*, 2008:6, Institute for Labour Market Policy Evaluation, Uppsala.

Donald, S. G., Y.-C. Hsuz, and R. P. Lieli (2011). Testing the unconfoundedness assumption via inverse probability weighted estimators of (l)att. Working Paper.

Fredriksson, P. and P. Johansson (2008). Dynamic treatment assignment – the consequences for evaluations using observational data. *Journal of Business and Economic and Statistics 26*, 435–445.

Frölich, M. (2007). Nonparametric iv estimation of local average treatment effects with covariates. *Journal of Econometrics 139*, 35–75.

Hahn, J., P. Todd, W. van der Klaaw, and W. Todd, Petra Van der Klaauw (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica 69*, 201–209.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics 86*, 4–29.

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

Johansson, P. (2008). The importance of employer contacts: Evidence based on selection on observables and internal replication. *Labour Economics 15*, 350–369.

Johansson, P. and S. Martinson (2000). Det nationella it-programmet - en slutrapport om swit. *Forskningsrapporter*, 2000:8, Institute for Labour Market Policy Evaluation, Uppsala.

Lauritzen, S. (1996). *Graphical Models*. Oxford: Oxford University Press.

Lee, M.-J. (2005). *Micro-econometrics for policy, program and treatment effects*. Oxford: Oxford University Press.

Neyman, J. (1923). Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych X*, 1–51. In Polish, English translation by D. Dabrowska and T. Speed in *Statistical Science*, **5**, 465–472, 1990.

Pearl, J. (2000). *Causality: Models, Reasoning and Inference*. Cambridge: Cambridge University Press.

Pearl, J. (2009). *Causality, Second edition*. Cambridge: Cambridge University Press.

Rivers, D. and H. Vuong (1988). Limited information estimators and exogeneity tests for simultaneous probit models. *Journal of Econometrics 39*, 347–366.

Rosenbaum, P. R. (1987). The role of a second control group in an observational study (with discussion). *Statistical Science 2*, 292–316.

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*, 41–55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology 66*, 688–701.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica 50*, 1–25.

Wooldridge, J. (2002). *Econometric Analysis of Cross Section and Panel Data.* Cambridge: MIT Press.