Testing the Unconfoundedness Assumption via Inverse Probability Weighted Estimators of (L)ATT *

Stephen G. Donald† Yu-Chin Hsu‡ Robert P. Lieli§

December 16, 2012

Abstract

We propose inverse probability weighted estimators for the the local average treatment effect (LATE) and the local average treatment effect for the treated (LATT) under instrumental variable assumptions with covariates. We show that these estimators are asymptotically normal and efficient. When the (binary) instrument satisfies one-sided non-compliance, we propose a Durbin-Wu-Hausman-type test of whether treatment assignment is unconfounded conditional on some observables. The test is based on the fact that under one-sided non-compliance LATT coincides with the average treatment effect for the treated (ATT). We conduct Monte Carlo simulations to demonstrate, among other things, that part of the theoretical efficiency gain afforded by unconfoundedness in estimating ATT survives pre-testing. We illustrate the practical implementation of the test on data from training programs administered under the Job Training Partnership Act.

JEL codes: C12, C13, C14

Keywords: local average treatment effect, instrumental variables, unconfoundedness, inverse probability weighted estimation, nonparametric estimation

*We thank Jason Abrevaya, Yu-Wei Hsieh, Chung-Ming Kuan, Blaise Melly, Chris Taber, Ed Vytlacil, the editors and three anonymous referees for useful comments. All errors are our responsibility.

†Department of Economics, University of Texas, Austin, donald@eco.utexas.edu.

‡Institute of Economics, Academia Sinica, ychsu@econ.sinica.edu.tw

§Department of Economics, Central European University, Budapest and the National Bank of Hungary, lielir@ceu.hu

1 Introduction

Nonparametric estimation of average treatment effects from observational data is typically undertaken under one of two types of identifying conditions. The unconfoundedness assumption, in its weaker form, postulates that treatment assignment is mean-independent of potential outcomes conditional on a vector of observed covariates. Nevertheless, even this requirement carries with it considerable identifying power; specifically, it identifies the average treatment effect (ATE) and the average treatment effect for the treated (ATT) without any additional modeling assumptions. On the other hand, if unobservable confounders exist then instrumental variables—related to the outcome only through changing the likelihood of treatment—are typically utilized to learn about treatment effects. Without further assumptions, the availability of an instrumental variable (IV) is however not sufficient to identify ATE or ATT. In general, the IV will only identify the local average treatment effect (LATE; Imbens and Angrist 1994) and the local average treatment effect for the treated (LATT; Frölich and Lechner 2010; Hong and Nekipelov 2010). If one specializes to binary instruments, as we do in this paper, then the LATE and LATT parameters correspond to the average treatment effect over specific subgroups of the population. These subgroups are however dependent on the choice of the instrument and are generally unobservable. Partly for these reasons a number of authors have called into question the usefulness of LATE for program evaluation (Heckman 1997; Heckman and Urzúa 2010; Deaton 2009). In most such settings ATE and ATT are more natural and practically relevant parameters of interest—provided that they can be credibly identified and accurately estimated. (In fairness, some of the criticism in Deaton (2009) goes beyond LATE, and also applies to ATE/ATT as a parameter of interest. See Imbens (2010) for a response to Deaton (2009).)

When using instrumental variables, empirical researchers are often called upon to tell a "story" to justify their validity. As pointed out by Abadie (2003) and Frölich (2007), it is often easier to argue that the relevant IV conditions hold if conditioning on a vector of observable covariates is also allowed. In particular, Frölich (2007) shows that in this scenario LATE is still nonparametrically identified and proposes efficient estimators, based on nonparametric imputation and matching, for this quantity. Given the possible need to condition on a vector of observables to justify the IV assumptions, it is natural to ask whether treatment assignment itself might be unconfounded conditional on the same (or maybe a larger or smaller) vector of covariates. In this paper we propose a formal test of this hypothesis that relies on the availability of a specific kind of binary instrument for

which LATT=ATT (so that the latter parameter is also identified). Establishing unconfoundedness under these conditions still offers at least two benefits: (i) it enables the estimation of an additional parameter of interest (namely, ATE) and (ii) it potentially allows for more efficient estimation of ATT than IV methods (we will argue this point in more detail later). To our knowledge this is the first test in the literature aimed at this task.

More specifically, the contributions of this work are twofold. First, given a (conditionally) valid binary instrument, we propose alternative nonparametric IV estimators of LATE and LATT. These estimators rely on weighting by the inverse of the estimated propensity score and are computed as the ratio of two estimators that are of the form proposed by Hirano et al. (2003), henceforth HIR. While Frölich (2007) conjectures in passing that such an estimator of LATE should be efficient, he does not provide a proof. We fill this (admittedly small) gap in the literature and formally establish the first order asymptotic equivalence of our LATE estimator and Frölich's imputation/matching-based estimators. We also demonstrate that our LATT estimator is asymptotically efficient, i.e. first-order equivalent to that of Hong and Nekipelov (2010).

More importantly, we propose a Durbin-Wu-Hausman-type test for the unconfoundedness assumption. On the one hand, if a binary instrument satisfying "one-sided non-compliance" (e.g., Frölich and Melly 2008a) is available, then the LATT parameter associated with that instrument coincides with ATT, and is consistently estimable using the estimator we proposed. (Whether one-sided non-compliance holds is verifiable from the data.) On the other hand, if treatment assignment is unconfounded given a vector of covariates, ATT can also be consistently estimated using the HIR estimator. If the unconfondedness assumption does not hold, then the HIR estimator will generally converge to a different limit. Thus, the unconfoundedness assumption can be tested by comparing our estimator of LATT with HIR's estimator of ATT. Of course, if the validity of the instrument itself is questionable, then the test should be more carefully interpreted as a joint test of the IV conditions and the unconfoundedness assumption. We use a battery of Monte Carlo simulations to explore in detail the finite sample properties of our IV estimator and the proposed test statistic. We also provide an application to illustrate how to implement and interpret the test in practice. We use the data set from Abadie et al. (2002) on training programs administered under the Job Training Partnership Act (JTPA) in the U.S. and highlight a difference between the self-selection process of men vs. women into these programs.

The rest of the paper is organized as follows. In Section 2 we present a (standard) framework

for defining and identifying causal effects nonparametrically. In Section 3 we propose estimators for LATE and LATT, and describe their asymptotic properties. The test for the unconfoundedness assumption is described in Section 4, along with its implications. A rich set of Monte Carlo results are presented in Section 5, and the empirical application is given in Section 6. Section 7 summarizes and concludes. The most important proofs are collected in a technical appendix.

2 The basic framework and identification results

The following IV framework, augmented by covariates, is now standard in the treatment effect literature; see, e.g., Abadie (2003) or Frölich (2007) for a more detailed exposition. For each population unit (individual) one can observe the value of a binary instrument $Z \in \{0,1\}$ and a vector of covariates $X \in \mathbb{R}^k$. For $Z = z$, the random variable $D(z) \in \{0,1\}$ specifies individuals' potential treatment status with $D(z) = 1$ corresponding to treatment and $D(z) = 0$ to no treatment. The actually observed treatment status is then given by $D \equiv D(Z) = D(1)Z + D(0)(1 - Z)$. Similarly, the random variable $Y(z,d)$ denotes the potential outcomes in the population that would obtain if one were to set $Z = z$ and $D = d$ exogenously. The following assumptions, taken from Abadie (2003) and Frölich (2007) with some modifications, describe the relationships between the variables defined above and justify Z being referred to as an instrument:

ASSUMPTION 1 Let $V = \big(Y(0,0), Y(0,1), Y(1,0), Y(1,1), D(1), D(0)\big)'$.
(i) (Moments): $E(V'V) < \infty$.
(ii) (Instrument assignment): There exists a subset X_1 of X such that $E(V|Z, X_1) = E(V|X_1)$ and $E(VV'|Z, X_1) = E(VV'|X_1)$ a.s. in X_1.
(iii) (Exclusion of the instrument): $P[Y(1,d) = Y(0,d)] = 1$ for $d \in \{0,1\}$.
(iv) (First stage): $P[D(1) = 1] > P[D(0) = 1]$ and $0 < P(Z = 1|X_1) < 1$ for X_1 defined in (ii).
(v) (Monotonicity): $P[D(1) \geq D(0)] = 1$.

Assumption 1(i) ensures the existence of the moments we will work with. Part (ii) states that, conditional on X_1, the instrument is exogenous with respect to the first and second moments of the potential outcome and treatment status variables. This is satisfied, for example, if the value of the instrument is completely randomly assigned or the instrument assignment is independent of V conditional on X_1. Nevertheless, part (ii) is weaker than the full conditional independence assumed in Abadie (2003) and Frölich (2007), but is still sufficient for identifying LATE and LATT.

Part (iii) precludes the instrument from having a direct effect on potential outcomes. Part (iv) postulates that the instrument is (positively) related to the probability of being treated and implies that the distributions $X_1|Z=0$ and $X_1|Z=1$ have common support. Finally, the monotonicity of $D(z)$ in z, required in part (v), allows for three different types of population units with nonzero mass: compliers $[D(0)=0, D(1)=1]$, always takers $[D(0)=1, D(1)=1]$ and never takers $[D(0)=0, D(1)=0]$ (cf. Imbens and Angrist 1994). Of these, compliers are actually required to have positive mass—part (iv) rules out $P[D(1)=D(0)]=1$. In light of these assumptions it is customary to think of Z as a variable that indicates whether an exogenous incentive to obtain treatment is present or as a variable signaling "intention to treat".

Given the exclusion restriction in part (iii), one can simplify the definition of the potential outcome variables as $Y(d) \equiv Y(1,d) = Y(0,d)$, $d=0,1$. The actually observed outcomes are then given by $Y \equiv Y(D) = Y(1)D + Y(0)(1-D)$. The LATE ($\equiv \tau$) and LATT ($\equiv \tau_t$) parameters associated with the instrument Z are defined as

$$\begin{aligned} \tau &\equiv E[Y(1)-Y(0) \mid D(1)=1, D(0)=0], \\ \tau_t &\equiv E[Y(1)-Y(0) \mid D(1)=1, D(0)=0, D=1]. \end{aligned}$$

LATE, originally due to Imbens and Angrist (1994), is the average treatment effect in the complier subpopulation. The LATT parameter was considered, for example, by Frölich and Lechner (2010) and Hong and Nekipelov (2010). LATT is the average treatment effect among those compliers who actually receive the treatment. Of course, in the subpopulation of compliers the condition $D=1$ is equivalent to $Z=1$, i.e. LATT can also be written as $E[Y(1)-Y(0) \mid D(1)=1, D(0)=0, Z=1]$. In particular, if Z is an instrument that satisfies Assumption 1 unconditionally (say Z is assigned completely at random), then LATT coincides with LATE. Our interest in LATT is motivated mainly by the fact that it can serve as a bridge between the IV assumptions and unconfoundedness (this connection will be developed shortly).

Under Assumption 1 one can also interpret LATE/LATT as the ATE/ATT of Z on Y divided by the ATE/ATT of Z on D. More formally, define $W(z) \equiv D(z)Y(1) + (1-D(z))Y(0)$ and $W \equiv W(Z) = ZW(1) + (1-Z)W(0)$. It is easy to verify that $W = DY(1) + (1-D)Y(0) = Y$ and

$$\tau = E[W(1)-W(0)]/E[D(1)-D(0)] \tag{1}$$

$$\tau_t = E[W(1)-W(0) \mid Z=1]/E[D(1)-D(0) \mid Z=1]. \tag{2}$$

The quantities on the rhs of (1) and (2) are nonparametrically identified from the joint distribution

of the observables (Y, D, Z, X_1). We denote the conditional probability $P(Z = 1 \mid X_1)$ by $q(X_1)$ and refer to it as the propensity score. Under Assumption 1, the following identification results are implied, for example, by Theorem 3.1 in Abadie (2003):

$$E[W(1) - W(0)] = E\left[\frac{ZY}{q(X_1)} - \frac{(1-Z)Y}{1-q(X_1)}\right] \equiv \Delta \tag{3}$$

$$E[D(1) - D(0)] = E\left[\frac{ZD}{q(X_1)} - \frac{(1-Z)D}{1-q(X_1)}\right] \equiv \Gamma \tag{4}$$

$$E[W(1) - W(0)|Z = 1] = E\left[q(X_1)\left(\frac{ZY}{q(X_1)} - \frac{(1-Z)Y}{1-q(X_1)}\right)\right] \Big/ E[q(X_1)] \equiv \Delta_t \tag{5}$$

$$E[D(1) - D(0)|Z = 1] = E\left[q(X_1)\left(\frac{ZD}{q(X_1)} - \frac{(1-Z)D}{1-q(X_1)}\right)\right] \Big/ E[q(X_1)] \equiv \Gamma_t. \tag{6}$$

That is, $\tau = \Delta/\Gamma$ and $\tau_t = \Delta_t/\Gamma_t$.

The unconfoundedness assumption, introduced by Rosenbaum and Rubin (1983), is also known in the literature as selection-on-observables, conditional independence, or ignorability. We say that treatment assignment is unconfounded conditional on a subset X_2 of the vector X if

ASSUMPTION 2 (Unconfoundedness): $Y(1)$ and $Y(0)$ are mean-independent of D conditional on X_2, i.e. $E[Y(d)|D, X_2] = E[Y(d)|X_2]$, $d \in \{0, 1\}$.

Assumption 2 is stronger than Assumption 1 in the sense that it rules out systematic selection to treatment based on unobservable factors and permits nonparametric identification of ATE $= E[Y(1) - Y(0)]$ and ATT $= E[Y(1) - Y(0)|D = 1]$. For example, ATE is identified by an expression analogous to (3): replace Z with D and $q(X_1)$ with $p(X_2) = P(D = 1|X_2)$. Similarly, ATT is identified by an expression analogous to (4): make the same substitutions as above. (The condition $E[Y(0)|D, X_2] = E[Y(0)|X_2]$ is actually sufficient for identifying ATT.)

As mentioned above, ATE and ATT are often of more interest to decision makers than local treatment effects, but are not generally identified under Assumption 1 alone. A partial exception is when the instrument Z satisfies a strengthening of the monotonicity property called one-sided non-compliance (see, e.g., Frölich and Melly 2008a):

ASSUMPTION 3 (One-sided non-compliance): $P[D(0) = 0] = 1$.

As pointed out by a referee, Assumption 3 can be traced back to (at least) Bloom (1984), who estimated the effect of a program in the presence of "no shows". The stated condition means that those individuals for whom $Z = 0$ are excluded from the treatment group, while those for whom $Z = 1$ generally have the option to accept or decline treatment (e.g., Z might represent eligibility to

receive treatment). Hence, there are no always-takers; non-compliance with the intention-to-treat variable Z is only possible when $Z = 1$. More formally, for such an instrument $D = ZD(1)$, and so $D = 1$ implies $D(1) = 1$ (the treated are a subset of the compliers). Therefore,

$$\begin{aligned}\text{LATT} =& E[Y(1) - Y(0) \mid D(1) = 1, D(0) = 0, D = 1] \\ =& E[Y(1) - Y(0) \mid D(1) = 1, D = 1] \\ =& E[Y(1) - Y(0) \mid D = 1] = \text{ATT}. \end{aligned} \tag{7}$$

Thus, under one-sided non-compliance, ATT=LATT. The ATE parameter, on the other hand, remains generally unidentified under Assumptions 1 and 3 alone.

In Section 4 we will show how one can test Assumption 2 when a binary instrument, valid conditional on X_1 and satisfying one-sided non-compliance, is available. Frölich and Lechner (2010) also consider some consequences for identification of the IV assumption and unconfoundedness holding simultaneously (without one-sided non-compliance), but they do not discuss estimation by inverse probability weighting, propose a test, or draw out implications for efficiency.

3 The estimators and their asymptotic properties

3.1 Inverse propensity weighted estimators of LATE and LATT

Let $\{(Y_i, D_i, Z_i, X_{1i})\}_{i=1}^n$ denote a random sample of observations on (Y, D, Z, X_1). The proposed inverse probability weighted (IPW) estimators for τ and τ_t are based on sample analog expressions for (3) through (6):

$$\begin{aligned}\hat{\tau} &= \sum_{i=1}^n \left\{ \frac{Z_i Y_i}{\hat{q}(X_{1i})} - \frac{(1-Z_i)Y_i}{1-\hat{q}(X_{1i})} \right\} \Big/ \sum_{i=1}^n \left\{ \frac{Z_i D_i}{\hat{q}(X_{1i})} - \frac{(1-Z_i)D_i}{1-\hat{q}(X_{1i})} \right\}, \\ \hat{\tau}_t &= \sum_{i=1}^n \hat{q}(X_{1i}) \left\{ \frac{Z_i Y_i}{\hat{q}(X_{1i})} - \frac{(1-Z_i)Y_i}{1-\hat{q}(X_{1i})} \right\} \Big/ \sum_{i=1}^n \hat{q}(X_{1i}) \left\{ \frac{Z_i D_i}{\hat{q}(X_{1i})} - \frac{(1-Z_i)D_i}{1-\hat{q}(X_{1i})} \right\}, \end{aligned}$$

where $\hat{q}(\cdot)$ is a suitable nonparametric estimator of the the propensity score function. If there are no covariates in the model, then both $\hat{\tau}$ and $\hat{\tau}_t$ reduce to

$$\hat{\tau} = \hat{\tau}_t = \left\{ \frac{\sum_{i=1}^n Z_i Y_i}{\sum_{i=1}^n Z_i} - \frac{\sum_{i=1}^n (1-Z_i)Y_i}{\sum_{i=1}^n (1-Z_i)} \right\} \Big/ \left\{ \frac{\sum_{i=1}^n Z_i D_i}{\sum_{i=1}^n Z_i} - \frac{\sum_{i=1}^n (1-Z_i)D_i}{\sum_{i=1}^n (1-Z_i)} \right\},$$

which is the LATE estimator developed in Imbens and Angrist (1994) and is also known as the Wald estimator, after Wald (1940).

Following HIR, we use the series logit estimator (SLE) to estimate $q(\cdot)$. The first order asymptotic results presented in Section 3.2 do not depend critically on this choice—the same conclusions could be obtained under similar conditions if other suitable estimators of $q(\cdot)$ were used instead. For example, Ichimura and Linton (2005) use local polynomial regression. We opt for the SLE for three reasons: (i) it is automatically bounded between zero and one; (ii) it allows for a more unified treatment of continuous and discrete covariates in practice; (iii) the curse of dimensionality affects the implementability of the SLE less severely. (We do not say that the SLE is immune to dimensionality; however, when X_1 and the order of the local polynomial are both large, one either needs a very large bandwidth or an astronomical number of observations just to be able to compute each $\hat{q}(X_{1i})$, especially if a kernel with bounded support is used. A sufficiently restricted version of the SLE is always easy to compute.)

We implement the SLE using power series. Let $\lambda = (\lambda_1, ..., \lambda_r)' \in \mathbb{Z}_+^r$ be a r-dimensional vector of non-negative integers and define a norm for λ as $|\lambda| = \sum_{j=1}^r \lambda_j$. Let $\{\lambda(k)\}_{k=1}^\infty$ be a sequence including all distinct $\lambda \in \mathbb{Z}_+^r$ such that $|\lambda(k)|$ is non-decreasing in k and for $x_1 \in \mathbb{R}^r$, let $x_1^\lambda = \prod_{j=1}^r x_{1j}^{\lambda_j}$. For any integer K, define $R^K(x_1) = (x_1^{\lambda(1)}, ..., x_1^{\lambda(K)})'$ as a vector of power functions. Let the $\Lambda(a) = \exp(a)/(1+\exp(a))$ be the logistic CDF. The SLE for $q(X_{1i})$ is defined as $\hat{q}(x_1) = \Lambda\big(R^K(x_1)'\hat{\pi}_K\big)$ where

$$\hat{\pi}_K = \arg\max_{\pi_K \in \mathbb{R}^K} \sum_{i=1}^n \Big(Z_i \cdot \ln \Lambda\big(R^K(X_{1i})'\pi_K\big) + (1 - Z_i) \cdot \ln\big(1 - \Lambda\big(R^K(X_{1i})'\pi_K\big)\big)\Big).$$

The asymptotic properties of $\hat{q}(x_1)$ are discussed in Appendix A of HIR.

3.2 First order asymptotic results

We now state conditions under which $\hat{\tau}$ and $\hat{\tau}_t$ are $\sqrt{n}$-consistent, asymptotically normal and efficient.

ASSUMPTION 4 (Distribution of X_1): (i) The distribution of the r-dimensional vector X_1 is absolutely continuous with probability density $f(x_1)$; (ii) the support of X_1, denoted $\mathcal{X}_1$, is a Cartesian product of compact intervals; (iii) $f(x_1)$ is twice continuously differentiable, bounded above, and bounded away from 0 on $\mathcal{X}_1$.

Though standard in the literature, this assumption is restrictive in that it rules out discrete covariates. This is mostly for expositional convenience; after stating our formal result, we will discuss how to incorporate discrete variables into the analysis.

Next we impose restrictions on various conditional moments of Y, D and Z. We define $m_z(x_1) = E[Y \mid X_1 = x_1, Z = z]$ and $\mu_z(x_1) = E[D \mid X_1 = x_1, Z = z]$. Then:

ASSUMPTION 5 (Conditional Moments of Y and D): $m_z(x_1)$ and $\mu_z(x_1)$, are continuously differentiable over $\mathcal{X}_1$ for $z = 0, 1$.

ASSUMPTION 6 (Propensity Score): (i) $q(x_1)$ is continuously differentiable of order $\bar{q} \geq 7 \cdot r$; (ii) $q(x_1)$ is bounded away from zero and one on $\mathcal{X}_1$.

The last assumption specifies the estimator used for the propensity score function.

ASSUMPTION 7 (Propensity Score Estimator): The propensity score function is estimated by SLE with a power series satisfying $K = a \cdot n^{\nu}$ for some $1/(4(\bar{q}/r - 1)) < \nu < 1/9$ and $a > 0$.

The first-order asymptotic properties of $\hat{\tau}$ and $\hat{\tau}_t$ are stated in the following theorem.

Theorem 1 *(Asymptotic properties of $\hat{\tau}$ and $\hat{\tau}_t$): Suppose that Assumption 1 and Assumptions 4 through 7 are satisfied. Then:*

(a) $\sqrt{n}(\hat{\tau} - \tau) \xrightarrow{d} \mathcal{N}(0, \mathcal{V})$ *and* $\sqrt{n}(\hat{\tau}_t - \tau_t) \xrightarrow{d} \mathcal{N}(0, \mathcal{V}_t)$, *where* $\mathcal{V} = E[\psi^2(Y, D, Z, X_1)]$ *and* $\mathcal{V}_t = E[\psi_t^2(Y, D, Z, X_1)]$ *with the functions* ψ *and* ψ_t *given by*

$$\psi(y, d, z, x_1) = \frac{1}{\Gamma}\Bigg\{\frac{z\big[y - m_1(x_1) - \tau(d - \mu_1(x_1))\big]}{q(x_1)}$$
$$-\frac{(1-z)\big[y - m_0(x_1) - \tau(d - \mu_0(x_1))\big]}{1 - q(x_1)} + m_1(x_1) - m_0(x_1) - \tau[\mu_1(x_1) - \mu_0(x_1)]\Bigg\},$$
$$\psi_t(y, d, z, x_1) = \frac{q(x_1)}{E(Z)\Gamma_t}\Bigg\{\frac{z\big[y - m_1(x_1) - \tau_t(d - \mu_1(x_1))\big]}{q(x_1)}$$
$$-\frac{(1-z)\big[y - m_0(x_1) - \tau_t(d - \mu_0(x_1))\big]}{1 - q(x_1)} + \frac{z\big[m_1(x_1) - m_0(x_1) - \tau_t(\mu_1(x_1) - \mu_0(x_1))\big]}{q(x_1)}\Bigg\};$$

(b) $\mathcal{V}$ *is the semiparametric efficiency bound for LATE with or without the knowledge of* $q(x_1)$*;*

(c) $\mathcal{V}_t$ *the semiparametric efficiency bound for LATT without the knowledge of* $q(x_1)$*.*

Comments 1. The result on $\hat{\tau}$ is analogous to Theorem 1 of HIR; the result on $\hat{\tau}_t$ is analogous to Theorem 5 of HIR. Theorem 1 shows directly that the IPW estimators of LATE and LATT presented in this paper are first order asymptotically equivalent to the matching/imputation based estimators developed by Frölich (2007) and Hong and Nekipelov (2010).

2. Theorem 1 follows from the fact that, under the conditions stated, $\hat{\tau}$ and $\hat{\tau}_t$ can be expressed as asymptotically linear with influence functions ψ and ψ_t, respectively:

$$\sqrt{n}(\hat{\tau} - \tau) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(Y_i, D_i, Z_i, X_{1i}) + o_p(1), \tag{8}$$

$$\sqrt{n}(\hat{\tau}_t - \tau_t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_t(Y_i, D_i, Z_i, X_{1i}) + o_p(1). \tag{9}$$

These representations are developed in Appendix A, along with the rest of the proof.

3. To use Theorem 1 for statistical inference, one needs consistent estimators for $\mathcal{V}$ and $\mathcal{V}_t$. Such estimators can be obtained by constructing (uniformly) consistent estimates for ψ and ψ_t and then averaging the squared estimates over the sample observations $\{(Y_i, D_i, Z_i, X_{1i})\}_{i=1}^n$. To be more specific, let $\hat{m}_1(x_1)$ and $\hat{m}_0(x_1)$ be the series estimators for $m_1(x_1)$ and $m_0(x_1)$:

$$\begin{aligned}
\hat{m}_1(x_1) &= \left(\sum_{i=1}^{N} \frac{Y_i Z_i}{\hat{q}(X_{1i})} R^K(X_{1i}) \right)' \left(\sum_{i=1}^{N} R^K(X_{1i}) R^K(X_{1i})' \right)^{-1} R^K(x_1), \\
\hat{m}_0(x_1) &= \left(\sum_{i=1}^{N} \frac{Y_i (1 - Z_i)}{1 - \hat{q}(X_{1i})} R^K(X_{1i}) \right)' \left(\sum_{i=1}^{N} R^K(X_{1i}) R^K(X_{1i})' \right)^{-1} R^K(x_1),
\end{aligned} \tag{10}$$

where $R^K(x_1)$ is the same power series as in the SLE. As in HIR and Donald and Hsu (2012), it is true that $\sup_{x_1 \in \mathcal{X}_1} |\hat{m}_1(x_1) - m_1(x_1)| = o_p(1)$ and $\sup_{x_1 \in \mathcal{X}_1} |\hat{m}_0(x_1) - m_0(x_1)| = o_p(1)$. In addition, let $\hat{\mu}_1(x_1)$ and $\hat{\mu}_0(x_1)$ be defined by replacing Y_i with D_i in (10), and let $\widehat{E(Z)} = \sum_{i=1}^n \hat{q}(X_{1i})/n$. Construct the functions $\hat{\psi}(y, d, z, x_1)$ and $\hat{\psi}_t(y, d, z, x_1)$ by replacing $m_0(x_1)$, $m_1(x_1)$, $\mu_0(x_1)$, $\mu_1(x_1)$, $q(x_1)$, τ, τ_t, Γ, Γ_t, and $E(Z)$ with their estimators. (The estimators for Γ and Γ_t are the denominators of $\hat{\tau}$ and $\hat{\tau}_t$, respectively.) Finally, let

$$\widehat{\mathcal{V}} = \frac{1}{n} \sum_{i=1}^{n} \hat{\psi}(Y_i, D_i, Z_i, X_{1i})^2, \quad \widehat{\mathcal{V}}_t = \frac{1}{n} \sum_{i=1}^{n} \hat{\psi}_t(Y_i, D_i, Z_i, X_{1i})^2. \tag{11}$$

It is straightforward to show that $\widehat{\mathcal{V}} \xrightarrow{p} \mathcal{V}$ and $\widehat{\mathcal{V}}_t \xrightarrow{p} \mathcal{V}_t$.

4. In part (b), the semiparametric efficiency bound for τ is given by Frölich (2007) and Hong and Nekipelov (2010). In particular, the bounds are the same with or without the knowledge of the propensity score function $q(x_1)$. If the propensity score function is known, τ can also be consistently estimated by using the true propensity score throughout in the formula for $\hat{\tau}$. Denote this estimator by $\hat{\tau}^*$. Using Remark 2 of HIR, the asymptotic variance of $\sqrt{n}(\hat{\tau}^* - \tau)$ is $\mathcal{V}^* = E[(\psi^*(Y_i, D_i, Z_i, X_{1i}))^2]$ with

$$\psi^*(y, d, z, x_1) = \psi(y, d, z, x_1) + (z - q(x_1)) \Big(\frac{m_1(x_1) - \tau \mu_1(x_1)}{q(x_1)} + \frac{m_0(x_1) - \tau \mu_0(x_1)}{1 - q(x_1)} \Big).$$

It can be shown that $\mathcal{V} \leq \mathcal{V}^*$. Therefore, the IPW estimator for τ based on the true propensity score function is less efficient than that based on estimated propensity score function.

5. In part (c), the semiparametric efficiency bound for τ_t without knowledge of the propensity score function is derived in Hong and Nekipelov (2010). The efficiency bound for τ_t with knowledge of the propensity score function has not been given in the literature yet. However, by Corollary 1 and Theorem 5 of HIR, we expect that in this case the semiparametrically efficient IPW estimator for τ_t is given by

$$\hat{\tau}^*_{t,se} = \sum_{i=1}^n q(X_{1i}) \left\{ \frac{Z_i Y_i}{\hat{q}(X_{1i})} - \frac{(1-Z_i)Y_i}{1-\hat{q}(X_{1i})} \right\} \Big/ \sum_{i=1}^n q(X_{1i}) \left\{ \frac{Z_i D_i}{\hat{q}(X_{1i})} - \frac{(1-Z_i)D_i}{1-\hat{q}(X_{1i})} \right\}.$$

The formal proof of this claim is beyond the immediate scope of this paper. The asymptotic variance of $\sqrt{n}(\hat{\tau}^*_{t,se} - \tau)$ is $\mathcal{V}^*_{t,se} = E[(\psi^*_{t,se}(Y_i, D_i, Z_i, X_{1i}))^2]$ with

$$\psi^*_{t,se}(y,d,z,x_1) = \frac{q(x_1)}{E(Z)\Gamma_t} \Bigg\{ \frac{z\big[y - m_1(x_1) - \tau_t(d - \mu_1(x_1))\big]}{q(x_1)}$$
$$- \frac{(1-z)\big[y - m_0(x_1) - \tau_t(d - \mu_0(x_1))\big]}{1-q(x_1)} + m_1(x_1) - m_0(x_1) - \tau_t(\mu_1(x_1) - \mu_0(x_1)) \Bigg\}.$$

It is true that $\mathcal{V}^*_{t,se} \leq \mathcal{V}_t$, i.e., knowledge of the propensity score allows for more efficient estimation of τ_t. On the other hand, if the propensity score function is known, then τ_t can also be consistently estimated by using $q(X_{1i})$ throughout in the formula for $\hat{\tau}^*_{t,se}$. It follows from the discussion after Corollary 1 of HIR that $\hat{\tau}_t$ and the latter estimator cannot generally be ranked in terms of efficiency.

6. Suppose that X_1 contains discrete covariates whose possible values partition the population into p subpopulations. Let the random variable $S \in \{1, \ldots, p\}$ denote the subpopulation a given unit is drawn from. For each s one can define $q(\tilde{X}_1, s) = P(Z = 1 \mid \tilde{X}_1, S = s)$, where $\tilde{X}_1$ denotes the continuous components of X_1, and estimate this function by SLE on the corresponding subsample. Then LATE and LATT can be estimated in the usual way by using $\hat{q}(\tilde{X}_i, S_i)$ as the propensity score (this is equivalent to computing the weighted average of the subpopulation-specific LATE/LATT estimates). Under suitable modifications of Assumptions 4 through 7, the LATE estimator so defined possesses an influence function $\psi(y, d, z, \tilde{x}_1, s)$ that is isomorphic to $\psi(y, d, z, x_1)$; one simply replaces the functions $q(x_1)$, $m_z(x_1)$, $\mu_z(x_1)$ with their subpopulation-specific counterparts $q(\tilde{x}_1, s)$, $m_z(\tilde{x}_1, s) = E[Y \mid \tilde{X}_1 = \tilde{x}_1, S = s, Z = z]$, etc. See Abrevaya et al. (2012), Appendix D, for a formal derivation in a similar context.

7. The changes in the regularity conditions required by the presence of discrete variables are straightforward. For example, Assumption 4 needs to hold for the conditional distribution

$\tilde{X}_1 \mid S = s$ for any s. The functions $m_z(\tilde{x}_1, s)$, etc., need to be continuously differentiable in $\tilde{x}$ for any s. Finally, in Assumptions 6 and 7, r is to be redefined as the dimension of $\tilde{X}_1$.

8. It is easy to specify the SLE so that it implements the sample splitting estimator described in comment 6 above in a single step. Given a vector $R^K(\tilde{x}_1)$ of power functions in $\tilde{x}_1$, use $R^K(x_1) = (1_{\{s=1\}}R^K(\tilde{x}_1)', \ldots, 1_{\{s=p\}}R^K(\tilde{x}_1)')$ in the estimation, i.e., interact the powers of $\tilde{x}_1$ with each subpopulation dummy. However, if p is large, the number of observations available from some of the subpopulations can be very small (or zero) even for large n. The SLE is well suited for bridging over data-poor regions using functional form restrictions. E.g., one can use

$$R^K(x_1) = (1_{\{s=1\}}R^L(\tilde{x}_1)', \ldots, 1_{\{s=p\}}R^L(\tilde{x}_1)', \tilde{x}_1^{\lambda(L+1)}, \ldots, \tilde{x}_1^{\lambda(K)})',$$

for some $L < K$, i.e., one only lets lower order terms vary across subpopulations. Alternatively, suppose that $X_1 = (\tilde{X}_1', I_1, I_2)'$ where I_1 and I_2 are two indicators (so that $p = 4$). Then one may implement the SLE with $(R^K(\tilde{x}_1), R^K(\tilde{x}_1)I_1, R^K(\tilde{x}_1)I_2)$, but without $R^K(\tilde{x}_1)I_1I_2$. This constrains the attributes I_1 and I_2 to operate independently from each other in affecting the probability that $Z = 1$. Of course, the two types of restrictions can be combined. The asymptotic theory is unaffected if the restrictions are removed for n sufficiently large. Furthermore, as restricting the SLE can be thought of as a form of smoothing, results by Li et al. (2009) suggest that using restrictions relative to the "sample splitting" method can lead to small sample MSE gains in estimating τ and τ_t (unless of course the misspecification bias is too large).

4 Testing for unconfoundedness

4.1 The proposed test procedure

If treatment assignment is unconfounded conditional on a subset X_2 of X, then, under regularity conditions, one can consistently estimate ATT using the estimator proposed by HIR:

$$\hat{\beta}_t = \sum_{i=1}^{n} \hat{p}(X_{2i}) \left\{ \frac{D_i Y_i}{\hat{p}(X_{2i})} - \frac{(1-D_i)Y_i}{1-\hat{p}(X_{2i})} \right\} \Big/ \sum_{i=1}^{n} \hat{p}(X_{2i}),$$

where $\hat{p}(x_2)$ is the series logit estimator of $p(x_2) = P(D = 1|X_2 = x_2)$. More generally, let $\rho_d(x_2) = E[Y|D = d, X_2 = x_2]$ for $d = 0, 1$. Under regularity conditions, $\hat{\beta}_t$ converges in probability to $\beta_t \equiv E[Y|D = 1] - E[\rho_0(X_2)|D = 1]$. If the unconfoundedness assumption holds, then $\rho_0(X_2) = E[Y(0) \mid D = 1, X_2]$, and β_t reduces to ATT$= E[Y(1) - Y(0)|D = 1]$.

Given a binary instrument that is valid conditional on a subset X_1 of X, one-sided non-compliance implies ATT=LATT, and hence ATT can also be consistently estimated by $\hat{\tau}_t$. On the other hand, if the unconfoundedness assumption does not hold, then $\hat{\tau}_t$ is still consistent for ATT, but $\hat{\beta}_t$ is in general not consistent for ATT. Hence, we can test the unconfoundedness assumption (or at least a necessary condition of it) by comparing $\hat{\tau}_t$ with $\hat{\beta}_t$. In particular, let

$$\phi_t(y,d,x_2) = \frac{p(x_2)}{p}\left\{\frac{d(y-\rho_1(x_2))}{p(x_2)} - \frac{(1-d)(y-\rho_0(x_2))}{1-p(x_2)} + \frac{d(\rho_1(x_2)-\rho_0(x_2)-\beta_t)}{p(x_2)}\right\},$$

where $p = P(D=1)$. The asymptotic properties of the difference between $\hat{\tau}_t$ and $\hat{\beta}_t$ are summarized in the following theorem:

Theorem 2 *Suppose that Assumption 1 and Assumptions 4 through 7 are satisfied. If some additional regularity conditions stated by HIR hold, and* $\sigma^2 \equiv E[(\psi_t(Y,D,Z,X_1) - \phi_t(Y,D,X_2))^2] > 0$, *then* $\sqrt{n}[(\hat{\tau}_t - \hat{\beta}_t) - (\tau_t - \beta_t)] \xrightarrow{d} N(0,\sigma^2)$. *If Assumptions 2 and 3 also hold, then* $\tau_t = \beta_t$*=(L)ATT.*

The additional regularity conditions referred to in Theorem 2 restrict the distribution of X_2, impose smoothness of $p(x_2)$, etc. HIR show that the asymptotic linear representation of $\hat{\beta}_t$ is given by

$$\sqrt{n}(\hat{\beta}_t - \beta_t) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_t(Y_i,D_i,X_{2i}) + o_p(1).$$

Theorem 2 follows directly from this result and Theorem 1. Let $\widehat{\psi}_t(\cdot)$ and $\widehat{\phi}_t(\cdot)$ be (uniformly) consistent estimators of ψ_t and ϕ_t obtained, e.g., as in comment 3 after Theorem 1. A consistent estimator for σ^2 can then be constructed as

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\psi}_t(Y_i,D_i,Z_i,X_{1i}) - \widehat{\phi}_t(Y_i,D_i,X_{2i})\right)^2.$$

Thus, if one-sided non-compliance holds, one can use a simple z-test with the statistic $\sqrt{n}(\hat{\tau}_t - \hat{\beta}_t)/\hat{\sigma}$ to test unconfoundedness via the null hypothesis $H_0 : \tau_t = \beta_t$. Since the difference between τ_t and β_t can generally be of either sign, a two-sided test is appropriate.

For the z-test to "work", it is also required that $\sigma^2 > 0$. It is difficult to list all cases where $\sigma^2 = 0$, but here we give one case that is easy to verify in practice. The proof is available on request.

Lemma 1 *Suppose that Assumption 1, Assumptions 4 through 7, and some additional regularity conditions stated by HIR are satisfied. If* $Var(Y(0)) = 0$, *then* $\tau_t = \beta_t$ *and* $\sqrt{n}(\hat{\tau}_t - \hat{\beta}_t) = o_p(1)$.

Further comments 1. The result in Theorem 2 holds without unconfoundedness, providing consistency against violations for which $\beta_t \neq \tau_t$. Nevertheless, as suggested by Lemma 1, unconfoundedness might be violated even when H_0 holds. The condition $E[Y(0)|D, X_2] = E[Y(0)|X_2]$ is actually sufficient (and necessary) to identify and consistently estimate ATT. Therefore, our test will not have power against cases where $E[Y(0)|D, X_2] = E[Y(0)|X_2]$ but $E[Y(1)|D, X_2] \neq E[Y(1)|X_2]$.

2. The proposed test is quite flexible in that it does not place any restrictions on the relationship between X_1 and X_2. The two vectors can overlap, be disjoint, or one might be contained in the other. The particular case in which X_2 is empty corresponds to testing whether treatment assignment is completely random.

3. If the instrument is not entirely trusted (even with conditioning), then the interpretation of the test should be more conservative; namely, it should be regarded as a joint test of unconfoundedness and the IV conditions. A rejection in this case puts the researcher back to "square one" in that one cannot even be sure that LATE and (L)ATT are identified.

4. Our test generalizes to one-sided non-compliance of the form: $P[D(1) = 1] = 1$, i.e., where all units with $Z = 1$ will get treatment and only part of the units with $Z = 0$ can get treatment. To this end, define LATE for the non-treated as LATNT$\equiv \tau_{nt} \equiv E[Y(1)-Y(0)|D(1) = 1, D(0) = 0, D = 0]$ and ATE for the non-treated as ATNT$\equiv \beta_{nt} \equiv E[Y(1) - Y(0)|D = 0]$. Similarly to (7), we have LATNT=ATNT when $P[D(1) = 1] = 1$. We can estimate τ_{nt} by

$$\hat{\tau}_{nt} = \sum_{i=1}^{n}(1 - \hat{q}(X_{1i}))\left\{\frac{Z_i Y_i}{\hat{q}(X_{1i})} - \frac{(1 - Z_i)Y_i}{1 - \hat{q}(X_{1i})}\right\} \Big/ \sum_{i=1}^{n}(1 - \hat{q}(X_{1i}))\left\{\frac{Z_i D_i}{\hat{q}(X_{1i})} - \frac{(1 - Z_i)D_i}{1 - \hat{q}(X_{1i})}\right\}.$$

The corresponding estimator for β_{nt}, denoted $\hat{\beta}_{nt}$, has the same form as the numerator of $\hat{\tau}_{nt}$ with D_i replacing Y_i and $\hat{p}(X_{2i})$ replacing $\hat{q}(X_{1i})$. The logic of the test remains the same: (L)ATNT can be consistently estimated by $\hat{\tau}_{nt}$ as well as $\hat{\beta}_{nt}$ under the unconfoundedness assumption. However, if unconfoundedness does not hold, then $\hat{\tau}_{nt}$ is still consistent for ATNT, but $\hat{\beta}_{nt}$ is generally not. The technical details are similar to the previous case and are omitted.

5. If $P[D(0) = 0] = 1$ and $P[D(1) = 1] = 1$ both hold then $Z = D$ and instrument validity and unconfoundedness are one and the same. Furthermore, in this case $\sqrt{n}(\hat{\tau}_t - \hat{\beta}_t) = o_p(1)$.

4.2 The implications of unconfoundedness

What are the benefits of (potentially) having the unconfoundedness assumption at one's disposal in addition to IV conditions? An immediate one is that the ATE parameter also becomes identified

and can be consistently estimated, for example, by the IPW estimator proposed by HIR or by nonparametric imputation as in Hahn (1998).

A more subtle consequence has to do with the efficiency of $\hat{\beta}_t$ and $\hat{\tau}_t$ as estimators of ATT. If an instrument satisfying one-sided compliance is available, and the unconfoundedness assumption holds at the same time, then both estimators are consistent. Furthermore, the asymptotic variance of $\hat{\tau}_t$ attains the semiparametric efficiency bound that prevails under the IV conditions alone, and the asymptotic variance of $\hat{\beta}_t$ attains the corresponding bound that can be derived from the unconfoundedness assumption alone. The simple conjunction of these two identifying conditions does not generally permit an unambiguous ranking of the efficiency bounds even when $X_1 = X_2$. Nevertheless, by taking appropriate linear combinations of $\hat{\beta}_t$ and $\hat{\tau}_t$, one can obtain estimators that are more efficient than either of the two. This observation is based on the following elementary lemma:

Lemma 2 *Let A_0 and A_1 be two random variables with $\text{var}(A_0) < \infty$, $\text{var}(A_1) < \infty$ and $\text{var}(A_1 - A_0) > 0$. Define $A_a = (1-a)A_0 + aA_1$ for any $a \in \mathbb{R}$. Let $\bar{a} = \frac{\text{var}(A_0) - \text{cov}(A_0, A_1)}{\text{var}(A_1 - A_0)}$. Then: (a) $\text{var}(A_{\bar{a}}) \leq \text{var}(A_a)$ for all $a \in \mathbb{R}$; (b) $\text{var}(A_{\bar{a}}) < \text{var}(A_0)$ when $\bar{a} \neq 0$, i.e. $\text{var}(A_0) \neq \text{cov}(A_0, A_1)$; (c) $\text{var}(A_{\bar{a}}) < \text{var}(A_1)$ when $\bar{a} \neq 1$, i.e. $\text{var}(A_1) \neq \text{cov}(A_0, A_1)$.*

Let $\hat{\beta}_t(a) = (1-a)\hat{\beta}_t + a\hat{\tau}_t$ and $\mathcal{V}_t(a) = \text{var}[(1-a)\phi_{ti} + a\psi_{ti}]$, where $\psi_{ti} = \psi_t(Y_i, D_i, Z_i, X_{1i})$ and $\phi_{ti} = \phi_t(Y_i, D_i, X_{2i})$. Lemma 2 implies that for any $a \in \mathbb{R}$, $\hat{\beta}_t(a)$ is consistent for τ_t and is asymptotically normal with asymptotic variance $\mathcal{V}_t(a)$, i.e. $\sqrt{n}(\hat{\beta}_t(a) - \tau_t) \xrightarrow{d} \mathcal{N}(0, \mathcal{V}_t(a))$. The optimal weight $\bar{a}$ can be obtained as

$$\bar{a} = \frac{\text{var}(\phi_t) - \text{cov}(\phi_t, \psi_t)}{\text{var}(\phi_t) + \text{var}(\psi_t) - 2\text{cov}(\phi_t, \psi_t)},$$

so that $\mathcal{V}_t(\bar{a}) \leq \mathcal{V}_t(a)$ for all $a \in \mathbb{R}$. In other words, $\hat{\beta}_t(\bar{a})$ will be the most efficient estimator among all linear combinations of $\hat{\beta}_t$ and $\hat{\tau}_t$. Although $\bar{a}$ is unknown in general, it can be consistently estimated by

$$\hat{a} = \frac{\sum_{i=1}^n \widehat{\phi}_t(Y_i, D_i, Z_i, X_{1i})\big(\widehat{\phi}_t(Y_i, D_i, Z_i, X_{1i}) - \widehat{\psi}_t(Y_i, D_i, X_{2i})\big)}{\sum_{i=1}^n \big(\widehat{\phi}_t(Y_i, D_i, Z_i, X_{1i}) - \widehat{\psi}_t(Y_i, D_i, X_{2i})\big)^2}.$$

The asymptotic equivalence lemma, e.g., Lemma 3.7 of Wooldridge (2010), implies that $\sqrt{n}(\hat{\beta}_t(\hat{a}) - \tau_t)$ has the same asymptotic distribution as $\sqrt{n}(\hat{\beta}_t(\bar{a}) - \tau_t)$.

If $Var(\phi_t) = Cov(\phi_t, \psi_t)$, then $\bar{a} = 0$, which implies that $\hat{\beta}_t$ itself is more efficient than $\hat{\tau}_t$ (or any linear combination of the two). We give sufficient conditions for this result.

Theorem 3 *Suppose that Assumption 1 parts (i), (iii), (iv), (v) and Assumption 3 are satisfied, and let* $V = (Y(0), Y(1))$. *If, in addition,* $X_1 = X_2 = X$,

$$E(V \mid Z, D, X) = E(V \mid X) \quad \textit{and} \quad E(VV' \mid Z, D, X) = E(VV' \mid X), \tag{12}$$

then $\bar{a} = 0$.

The proof of Theorem 3 is provided in Appendix B. The conditions of Theorem 3 are stronger than those of Theorem 2. The latter theorem only requires that the IV assumption and unconfoundedness both hold at the same time, which in general does not imply the stronger *joint* mean-independence conditions given in (12). If the null of unconfoundedness is accepted due to (12) actually holding, then $\hat{\beta}_t$ itself is the most efficient estimator of ATT in the class $\{\hat{\beta}_t(a) : a \in \mathbb{R}\}$.

The theoretical results discussed in this subsection are qualified by the fact that in practice one needs to pre-test for unconfoundedness, while the construction of $\hat{\beta}_t(\bar{a})$ takes this assumption as given. Deciding whether or not to take a linear combination based on the outcome of a test will erode some of the theoretically possible efficiency gain when unconfoundedness does hold, and will introduce at least some bias through type 2 errors. (A related problem, the impact of a Hausman pre-test on subsequent hypothesis tests, was studied recently by Guggenberger 2010.) We will explore the effect of pre-testing in some detail through the Monte Carlo simulations presented in the next section.

5 Monte Carlo simulations

We employ a battery of simulations to gain insight into how accurately the asymptotic distribution given in Theorem 1 approximates the finite sample distribution of our LATE estimator in various scenarios and to gauge the size and power properties of the proposed test statistic. The scenarios examined differ in terms of the specification of the propensity score, the choice of the power series used in estimating it, and the trimming applied to the estimator. All these issues are known to be central for the finite sample properties of an IPW estimator. We also address the pre-testing problem raised at the end of Section 4; namely, we examine how much of the theoretical efficiency gain afforded by unconfoundedness is eroded by testing for it, and also the costs resulting from type 2 errors in situations when the assumption does not hold.

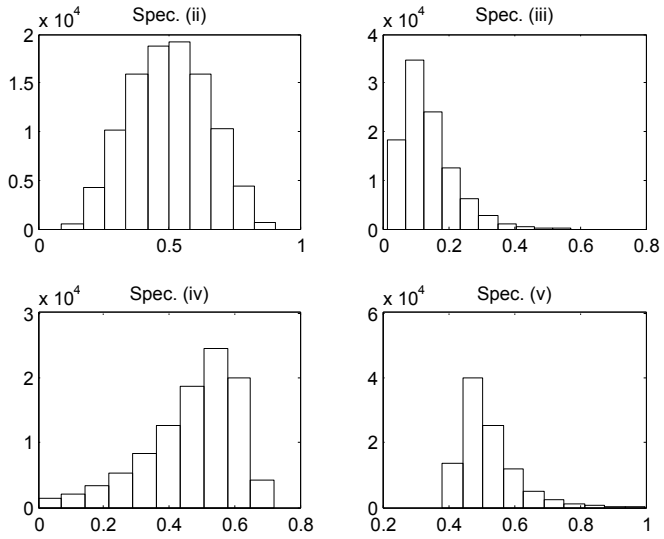

Figure 1: The distribution of $q(X)$

We use DGPs of the following form:

$$Y = (1 - D) \cdot (g(X) + \epsilon), \quad D = Z \cdot 1(b\epsilon + (1 - b)\nu > h(X)), \quad Z = 1(q(X) > U)$$

where $b \in [0, 1]$, $X = (W_1, \ldots, W_5)$, $g(X) = W_1 + \ldots + W_5$, $h(X) = W_4 - W_5$; the components of X and the unobserved errors ϵ, ν, U are mutually independent with

$$W_i \sim \text{unif}[0, 1], \;\; \epsilon \sim \text{unif}[-1, 1], \;\; \nu \sim \text{unif}[-1, 1], \;\; U \sim \text{unif}[0, 1].$$

We consider five different specifications for the propensity score function: (i) Constant: $q(X) = 0.6$; (ii) Linear 1: $q(X) = \Lambda(-2.5 + g(X))$; (iii) Linear 2: $q(X) = \Lambda(-4.5 + g(X))$; (iv) Rational expression 1: $q(X) = \Lambda(2 - 5/g(X))$; (v) Rational expression 2: $q(X) = \Lambda(-1 + 2.5/g(X))$. The distribution of the random variable $q(X)$ in the various cases is shown on Figure 1.

We choose X to be moderately high dimensional ($k = 5$), as in practice one would typically need a handful of covariates to make a credible case for unconfoundedness. Clearly, the DGP satisfies one-sided non-compliance as $D(0) = 0$. Finally, the value of the parameter $b \in [0, 1]$ governs whether the unconfoundedness assumption is satisfied. In particular, when $b = 0$ unconfondedness holds conditional on $X_2 = X$. Larger values of b correspond to more severe violations of the unconfoundedness assumption. The instrument Z is valid conditional on $X_1 = X$ for any b.

5.1 The finite sample distribution of the LATE estimator

In our first exercise we study the finite sample distribution of the LATE estimator $\hat{\tau}$. The DGP is designed so that the true value of the LATE parameter is independent of $q(\cdot)$ and is approximately equal to $\tau = -2.73$ for $b = 0.5$, the value chosen for this exercise.

Table 2 shows various statistics characterizing the finite sample distribution of $\hat{\tau}$ and its estimated standard error for $n = 250$, $n = 500$ and $n = 2500$. In particular, we report the bias of $\hat{\tau}$, the standard error of $T \equiv \sqrt{n}(\hat{\tau} - \tau)$, the mean estimate of this standard error based on comment 3 after Theorem 1, and the tail probabilities of the studentized estimator $S \equiv (\hat{\tau} - E\hat{\tau})/\widehat{s.e.}(\hat{\tau})$ associated with the critical values -1.645 and 1.645. The number of Monte Carlo repetitions is 5000.

For each specification of $q(\cdot)$, we consider a number of implementations of the SLE. We start with a constant model for the propensity score and then add linear, quadratic and cubic terms (all powers of W_i and all cross products up to the given order). We use the same power series to estimate all other nonparametric components of the influence function (used in estimating the standard error of $\hat{\tau}$). The choice of the power series in implementing the SLE is an important one; it mimics the choice of the smoothing parameter in kernel-based or local polynomial estimation. To our knowledge, there is no well-developed theory to guide the power series choice in finite samples (though Imbens et al. 2007 is a step in this direction); hence, a reasonable strategy in practice would involve examining the sensitivity of results to various specifications as is done in this simulation.

When using an IPW estimator in practice, the estimated probabilities are often trimmed to prevent them from getting too close to, or crossing, the boundaries of the [0,1] interval. Therefore, we also apply trimming to the raw estimates delivered by the SLE. The column "Trim." in Table 2 denotes the truncation applied to the estimated propensity scores. A value of $\gamma \in (0, 1/2)$ means that the propensity score estimate is forced to lie in the interval $[\gamma, 1-\gamma]$. We use $\gamma = 0.5\%$ (mild trimming) and, occasionally, $\gamma = 5\%$ (aggressive trimming).

Many aspects of the results displayed in Table 2 merit discussion.

First, looking at the simplest case when neither q and nor $\hat{q}$ depends on X, we see that even for $n = 250$, the bias of the LATE estimator is very small, its estimated standard error, too, is practically unbiased, and the distribution of the studentized estimator has tail probabilities close to standard normal. Even though the true propensity score does not depend on the covariates, one can achieve a substantial reduction in the standard error of the estimator by allowing $\hat{q}$ to be a

function of X, as suggested by Theorem 3 of Frölich and Melly (2008b). For example, when $\hat{q}(X)$ is linear, the standard error, for $n = 2500$, falls from about 3.21 to 2.42, roughly a 25% reduction. Nevertheless, we can also observe that if $\hat{q}(X)$ is very generously parameterized (here: quadratic), then in small samples the "noise" from estimating too many zeros can overpower most of this efficiency gain. Specifically, for $n = 250$ the standard error of the scaled estimator is almost back up to the no-covariate case (3.16 vs. 3.23). Still, the efficiency gains are recaptured for large n.

A second, perhaps a bit more subtle, point can be made about the standard error of $\hat{\tau}$ using the Linear 1 specification for $q(X)$. Here the linear SLE acts as a correctly specified parametric estimator while the estimated standard errors are computed under the assumption that q is non-parametrically estimated. Therefore, the estimated standard errors are downward-biased, reflecting the fact that even when the propensity score is known up to a finite dimensional parameter vector, it is more efficient to use a nonparametric estimator in constructing $\hat{\tau}$ as in Chen et al. (2008). Indeed, as the SLE adds quadratic and cubic terms, i.e., it starts 'acting' more as a nonparametric estimator, the bias vanishes from the estimated standard errors, provided that the sample size expands simultaneously ($n = 2500$). Furthermore, the asymptotic standard errors associated with the quadratic and cubic SLE (2.72 and 2.93, respectively) are lower than for the linear (3.11). In cases where the variance of $\hat{\tau}$ is underestimated, the studentized estimator tends to have more mass in its tails than the standard normal distribution (see, e.g., the results for the linear SLE).

Third, as best demonstrated by the Linear 2 model for the propensity score, the limit distribution provided in Theorem 1 can be a poor finite sample approximation when $q(X)$ gets close to zero or one with relatively high probability. This is especially true when the estimator for $q(X)$ is overspecified (quadratic or cubic). For $n = 250$ and $n = 500$, the bias of $\hat{\tau}$ ranges from moderate to severe and is exacerbated by more aggressive trimming of $\hat{q}$. For any series choice, the standard error of the LATE estimator is larger than in the Linear 1 case (the 0.5% vs. 5% trimming does not change the actual standard errors all that much). Furthermore, for $\hat{q}$ quadratic or cubic, the estimated standard errors are severely upward biased with mild trimming, and still very much biased, though in the opposite direction, with aggressive trimming. Increasing the sample size to $n = 2500$ of course lessens these problems, though judging from the tail probabilities, the standard normal can remain a rather crude approximation to the studentized estimator. E.g., for the cubic SLE with 0.5% trimming the standard error is grossly overestimated and there is evidence of skewness. On the other hand, for the linear and quadratic SLE the estimated asymptotic standard errors display

downward bias, presumably due to the "correct parametric specification" issue discussed in the second point above. Somewhat surprisingly, though, the actual standard errors are the smallest for the linear SLE; apparently, even for $n = 2500$, there is more than 'optimal' noise in the quadratic and cubic propensity score estimates.

Fourth, when the propensity score estimator is underspecified, $\hat{\tau}$ is an asymptotically biased estimator of LATE. (Here 'underspecified' refers to a misspecified model in the parametric sense or, in the context of series estimation, extending the power series too slowly as the sample size increases.) The bias is well seen in all cases in which the propensity score depends on X, but is estimated by a constant. The Rat. 1 and Rat. 2 models provide further illustration. Here, any fixed power series implementation of the SLE is misspecified if regarded as a parametric model, though the estimator provides an increasingly better approximation to $q(\cdot)$ as the power series expands. For the Rat. 1 model, the bias of $\hat{\tau}$ indeed decreases in magnitude as the SLE becomes more and more flexible, with the exception of $n = 250$. For Rat. 2, even the linear SLE removes the bias almost completely and not much is gained, even asymptotically, by using a more flexible estimator. For Rat. 1 there is noticeable asymptotic bias in estimating the standard error of $\hat{\tau}$, which would presumably disappear if the sample size and the power series both expanded further. Nevertheless, for both rational models the normal approximation to $\hat{\tau}$ works reasonably well in large samples across a range of implementations of the SLE.

Finally, the results as a whole show the sensitivity of $\hat{\tau}$ to the specification of the power series used in estimating the the propensity score $q(\cdot)$. If the power series has too few terms (or expands too slowly with the sample size), then $\hat{\tau}$ may be (asymptotically) biased. On the other hand, using too flexible a specification for a given sample size can cause $\hat{\tau}$ to have severe small sample bias and inflated variance, which is also estimated with bias. More aggressive trimming of the propensity score tends to increase the bias of $\hat{\tau}$ and reduce the bias of $\widehat{s.e.}(\hat{\tau})$, though to an uncertain degree.

5.2 Properties of the test and the pre-tested estimator

We first set $b = 0$ so that unconfoundedness holds for any specification of $q(X)$ conditional on $X_2 = X$. All tests are conducted at the 5% nominal significance level and with $X_1 = X_2 = X$, i.e., we drop the cases where $\hat{q}$ is constant. To further economize on space, we also drop the 5% truncation for the Rat. 1 specification. In each of the remaining cases we consider four estimators of (L)ATT: $\hat{\tau}_t$, $\hat{\beta}_t$, their combination $\hat{\beta}_t(\hat{a})$, and a pre-tested estimator, given by $\hat{\beta}_t(\hat{a})$ whenever the

test accepts unconfoundedness and $\hat{\tau}_t$ when it rejects it. Trimming is also applied to $\hat{p}(\cdot)$.

In Tables 3 and 4 we report, for each estimator, the raw bias, the standard deviation of $\sqrt{n}(\widehat{(L)ATT} - (L)ATT)$, the mean of the estimated standard deviation, and the mean squared error of $\sqrt{n}(\widehat{(L)ATT} - (L)ATT)$. We use a naive (but natural) estimator for the standard error of the pre-tested estimator; namely, we take the estimated standard error of either $\hat{\beta}_t(a)$ or τ_t, depending on which one is used. In addition, we report the actual rejection rates and the average weight across Monte Carlo cycles that the combined estimator assigns to $\hat{\tau}_t$ (the mean of $\hat{a}$).

Again, several aspects of the results are worth discussing.

First, there is adequate, though not perfect, asymptotic size control in all cases where the specification of the SLE is sufficiently flexible and there is no excessive trimming. The extent to which the 5% trimming can distort the size of the test in the Lin. 2 case is rather alarming; in the very least, this suggests that trimming should be gradually eliminated as the sample size increases.

Second, in almost all cases, the combined estimator has smaller standard errors in small samples than the HIR estimator $\hat{\beta}_t$, and the drop is especially large when $\hat{q}$ is overspecified. While this tends to be accompanied by an uptick in absolute bias, in almost all cases the combined estimator has the lowest finite sample MSE—the only exceptions come from the Linear 2 model with aggressive trimming. As the DGP satisfies the conditions of Theorem 3, the combined estimator puts less and less weight on $\hat{\tau}_t$ in larger samples and becomes equivalent to $\hat{\beta}_t$ unless trimming interferes.

Third, even though the pre-tested estimator has a higher MSE than $\hat{\beta}_t$ or the combined estimator, in almost all the cases this MSE is lower than that of $\hat{\tau}_t$. (Again, the only exceptions come in the Linear 2 case with 5% trimming for $n = 2500$, but here $\hat{\beta}_t$ itself has a higher MSE than $\hat{\tau}_t$.) Thus, while there is a price to pay for testing the validity of the unconfoundedness assumption, there is still a substantial gain relative to the case where one only has the IV estimator to fall back on. Of course, one would be better off taking unconfoundedness at face value when it actually holds. But as we will shortly see, there is a large cost in terms of bias if one happens to be wrong, and the power of the unconfoundedness test helps avoid paying this cost.

Fourth, the naive method described above underestimates the true standard error of the pre-tested estimator. We briefly examined a bootstrap estimator in a limited number of cases, and the results (not reported) appear upward biased. We do not consider these results conclusive as we took some shortcuts due to computational cost (to study this estimator one has to embed a bootstrap cycle inside a Monte Carlo cycle). We further note that the distribution of the pre-tested estimator

can show severe departures from normality such as multimodality or extremely high kurtosis.

We now present cases where unconfoundedness does not hold conditional on X. Specifically, we set $b = 0.5$ again, but also show some results for $b = 0.25$. We focus only on those cases from the previous exercise where size was asymptotically controlled, as power has questionable value otherwise. The results are displayed in Table 5 ($b = 0.5$) and Table 6 ($b = 0.25$).

Our first point is that the test appears consistent against these departures from the null—rejection rates approach unity as the sample size grows in all cases examined. Nevertheless, overspecifying the series estimators can seriously erode power in small samples; see the cubic SLE in Table 5 for q=Lin. 1, Rat. 1, Rat. 2. In fact, in these cases the test is not unbiased. A further odd consequence of overfitting is that power need not increase monotonically with n; see again the cubic SLE in Table 5 for q=Lin. 2.

Second, $\hat{\beta}_t$ (and hence the combined estimator) is rather severely biased both in small samples and asymptotically, though the bias is of course smaller for $b = 0.25$. Therefore, even though $\hat{\beta}_t$ generally has a lower standard error than $\hat{\tau}_t$, its MSE, in large enough samples, is substantially larger than that of $\hat{\tau}_t$. As the sample size grows, the pre-tested estimator behaves more and more similarly to $\hat{\tau}$, eventually also dominating $\hat{\beta}_t$ and $\hat{\beta}(\hat{a})$.

Third, in smaller samples the MSE of the pre-tested estimator is often larger than that of τ_t as the pre-tested estimator uses $\hat{\beta}_t(\hat{a})$ with positive probability, and $\hat{\beta}_t(\hat{a})$ is usually inferior to $\hat{\tau}_t$ due to its bias inherited mostly from $\hat{\beta}_t$. However, there are cases in which the increased bias of the combined estimator is more than offset by a reduction in variance so that MSE($\hat{\beta}_t(\hat{a})$) is lower than MSE($\hat{\tau}_t$) or MSE($\hat{\beta}_t$) or both. This happens mainly when $n = 250$ and $\hat{q}$ is overspecified, and of course more "easily" when b is smaller. In fact, all the cases shown in Table 6 were chosen specifically to display this effect, but see also the cubic SLE for q=Lin. 1, Lin. 2, Rat. 1, Rat. 2 in Table 5. As in these cases power tends to be (very) low, the pre-tested estimator preserves most of the MSE gain delivered by $\hat{\beta}_t(\hat{a})$ or might even improve on it slightly. This property of the combined estimator mitigates the cost of the type 2 errors made by the test.

6 Empirical illustration

We apply our method to estimate the impact of JTPA training programs on subsequent earnings and to test the unconfoundedness of the participation decision. We use the same data set as Abadie

et al. (2002), henceforth AAI, publicly available at

http://econ-www.mit.edu/faculty/angrist/data1/data/abangim02

As described by Bloom et al. (1997) and AAI, part of the JPTA program (the National JTPA study) involved collecting data specifically for purposes of evaluation. In some of the service delivery areas, between Nov. 1987 and Sept. 1989, randomly selected applicants were offered a job-related service (classroom training, on-the-job training, job search assistance, probationary employment, etc.) or were denied services and excluded from the program for 18 months (1 out of 3 on average).

Clearly, the random offer of services (Z) can be used, without further conditioning, as an instrument for evaluating the effect of actual program participation (D) on earnings (Y), measured as the sum of earnings in the 30 month period following the offer. About 36 percent of those with an offer chose not to participate; conversely, a small fraction of applicants, less than 0.5 percent, ended up participating despite the fact that they were turned away. Hence, Z satisfies one-sided non-compliance almost perfectly; the small number of observations violating this condition were dropped from the sample. (AAI also ignore this small group in interpreting their results.) The total number of observations is then 11,150; of these, 6,067 are female and 5,083 are male. We treat the two genders separately throughout.

The full set of AAI covariates (X) include

> "dummies for black and Hispanic applicants, a dummy for high-school graduates (including GED holders), dummies for married applicants, 5 age-group dummies, and dummies for AFDC receipt (for women) and whether the applicant worked at least 12 weeks in the 12 months preceding random assignment. Also included are dummies for the original recommended service strategy [...] and a dummy for whether earnings data are from the second follow-up survey." (AAI, p. 101)

See Table 1 of AAI for descriptive statistics. To illustrate the "sample splitting" method described in comment 6 after Theorem 1 we also construct a smaller set of controls with dummies for high-school education, minority status (black or hispanic), and whether the applicant is below age 30.

In Table 1 we present four sets of estimation/test results. In the first exercise we do not use any covariates in computing $\hat{\tau}_t$ and $\hat{\beta}_t$. The LATT estimator $\hat{\tau}_t$ is interpreted as follows. Take, e.g., the value 1916.4 for females. This means that female compliers who actually participated in the program (i.e., were assigned $Z = 1$), are estimated to increase their 30-month earnings by \$1916.4

Table 1: Treatment effect estimates and unconfoundedness test results

Subpop.	Obs.	$\hat{\tau}_t$	std($\hat{\tau}_t$)	$\hat{\beta}_t$	std($\hat{\beta}_t$)	std($\hat{\tau}_t - \hat{\beta}_t$)	Test-stat	p-val. (2-sided)
$X_1 = X_2 = \emptyset$, $\hat{q}, \hat{p}$=const.								
Males	5083	1716.0	(916.4)	4035.7	(557.3)	(740.7)	-3.132	0.002
Females	6067	1916.4	(547.8)	2146.7	(346.4)	(436.1)	-0.528	0.597
$X_1 = X_2$ =(BELOW30, MINORITY, HS), $\hat{q}, \hat{p}$=sample splitting; const. within subsample								
Males	5083	1805.9	(904.5)	3936.0	(554.8)	(731.8)	-2.911	0.004
Females	6067	1813.7	(545.3)	1912.4	(347.2)	(434.6)	-0.227	0.820
$X_1 = X_2$ =(BELOW30, MINORITY, HS), $\hat{q}, \hat{p}$=linear								
Males	5083	1751.2	(905.6)	3922.3	(555.0)	(732.3)	-2.965	0.003
Females	6067	1809.9	(544.9)	1910.2	(346.9)	(434.0)	-0.231	0.817
$X_1 = X_2$ =full set of AAI controls, $\hat{q}, \hat{p}$=linear								
Males	5083	1666.7	(879.9)	3642.8	(546.2)	(709.9)	-2.784	0.005
Females	6067	1919.2	(522.3)	2108.2	(337.0)	(417.1)	-0.453	0.650

Note: $\hat{\tau}_t$ is the IPW IV estimator of (L)ATT. $\hat{\beta}_t$ is the IPW estimator of ATT under unconfoundedness. All estimates are in U.S. dollars (ca. 1990). Numbers in parenthesis are standard errors.

on average. Since Z is randomly assigned, this number can also be interpreted as an estimate of LATE, i.e. the average effect among all compliers. Further, by one-sided non-compliance, \$1916.4 is also an estimate of the female ATT. As the difference between $\hat{\tau}_t$ and $\hat{\beta}_t = 2146.7$ is not statistically significant, the hypothesis of completely random *participation* cannot be rejected for females. In contrast, $\hat{\beta}_t$ for males is more than twice as large as $\hat{\tau}_t$, and the difference is highly significant. This suggests that self-selection into the program among men is based partly on factors systematically related to the potential outcomes.

In the next two exercises we set X_1 and X_2 equal to the restricted set of covariates. First we split the male and female samples by the eight possible configurations of the three indicators and estimate the propensity score by the subsample averages of Z; then we restrict the functional form to logit with a linear index. The two sets of results are similar both to each other and the results from the previous exercise. In particular, random participation is not rejected for females while it is still strongly rejected for males. There are factors related to the male participation decision as well as the potential outcomes that are not captured by the set of covariates used.

Finally, in the fourth exercise we use the full set of AAI covariates in a linear logit model. Compared with the no-covariate case, the estimated standard errors are slightly lower across the

board, but the change in the point estimates are still within a small fraction of them. Once again, the test does not reject uncounfoundedness for females but it does for males.

Since the hypothesis of random treatment participation cannot be rejected for females, $\hat{\beta}_t$ can also be interpreted as an estimate of ATE. In contrast, $\hat{\beta}_t$ is likely to be substantially biased as an estimate of male ATE. Furthermore, based on Section 4, one can take a weighted average of $\hat{\tau}_t$ and $\hat{\beta}_t$ to obtain a more efficient estimate of female ATE/ATT. As $\hat{a} \approx 0$ in all cases, the combined estimator is virtually the same as $\hat{\beta}_t$ and is not reported. Nevertheless, without testing for (and accepting) the unconfoundedness assumption, the only valid estimate of female ATT is $\hat{\tau}_t$, which has a much larger standard error than $\hat{\beta}_t$.

While the result on male vs. female self-selection is robust in this limited set of exercises, one would need to study the program design in more detail before jumping to conclusions about, say, behavioral differences. Understanding how the explicitly observed violations of one-sided non-compliance came about would be especially pertinent, and, as pointed out by a referee, the broader issue of control group substitution documented by Heckman et al. (2000) would also have to be taken into account. Furthermore, there are potentially relevant covariates (e.g., indicators of the service delivery area) not available in the AAI version of the data set. In short, the empirical results are best treated as illustrative or as a starting point for a more careful investigation.

7 Conclusion

Given a (conditionally) valid binary instrument, nonparametric estimators of LATE and LATT can be based on imputation or matching, as in Frölich (2007), or weighting by the estimated propensity score, as proposed in this paper. The two approaches are shown to be asymptotically equivalent; in particular, both types of estimators are $\sqrt{n}$-consistent and efficient.

When the available binary instrument satisfies one-sided non-compliance, the proposed estimator of LATT is compared with the ATT estimator of HIR to test the assumption that treatment assignment is unconfounded given a vector of observed covariates. To our knowledge, this is the first such test in the literature. Acceptance of uncofoundedness allows one to estimate ATE and improve on the asymptotic variance of the IV-based (L)ATT estimator. Simulations show that there are finite sample MSE gains even after the pre-testing effect is taken into account. An illustrative application of the test using JTPA data rejects unconfoundedness for males but not for females.

Appendix: Proofs

A. The proof of Theorem 1

We only show the proof for $\hat{\tau}$. The treatment of $\hat{\tau}_t$ is similar and is available upon request. In order to simplify notation, we set $X_1 = X$. Let

$$\widehat{\Delta} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{Z_i Y_i}{\hat{q}(X_i)} - \frac{(1-Z_i)Y_i}{1-\hat{q}(X_i)}\right\}, \qquad \widehat{\Gamma} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{Z_i D_i}{\hat{q}(X_i)} - \frac{(1-Z_i)D_i}{1-\hat{q}(X_i)}\right\}.$$

so that $\hat{\tau} = \widehat{\Delta}/\widehat{\Gamma}$. The asymptotic properties of $\hat{\Delta}$ and $\hat{\Gamma}$ are established in the following lemma.

Lemma 3 *Under the conditions of Theorem 1,* $\sqrt{n}(\widehat{\Delta}-\Delta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\delta(Y_i, D_i, Z_i, X_i) + o_p(1)$ *and* $\sqrt{n}(\widehat{\Gamma}-\Gamma) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\gamma(Y_i, D_i, Z_i, X_i) + o_p(1)$*, where*

$$\delta(Y_i, D_i, Z_i, X_i) = \frac{Z_i Y_i}{q(X_i)} - \frac{(1-Z_i)Y_i}{1-q(X_i)} - \Delta - \left(\frac{m_1(X_i)}{q(X_i)} + \frac{m_0(X_i)}{1-q(X_i)}\right)(Z_i - q(X_i))$$
$$\gamma(Y_i, D_i, Z_i, X_i) = \frac{Z_i D_i}{q(X_i)} - \frac{(1-Z_i)D_i}{1-q(X_i)} - \Gamma - \left(\frac{\mu_1(X_i)}{q(X_i)} + \frac{\mu_0(X_i)}{1-q(X_i)}\right)(Z_i - q(X_i)).$$

To make use of Lemma 3 we take a first order Taylor expansion of $\hat{\Delta}/\hat{\Gamma}$ around the point (Δ, Γ), yielding

$$\sqrt{n}(\hat{\tau}-\tau) = \sqrt{n}\left(\frac{\widehat{\Delta}}{\widehat{\Gamma}} - \frac{\Delta}{\Gamma}\right) = \frac{1}{\Gamma}\sqrt{n}(\widehat{\Delta}-\Delta) - \frac{\tau}{\Gamma}\sqrt{n}(\widehat{\Gamma}-\Gamma) + o_p(1). \tag{13}$$

Applying Lemma 3 to (13) gives (8). Under Assumption 1(i), we have $E[\psi(Y, D, Z, X)] = 0$ and $E[\psi^2(Y, D, Z, X)] < \infty$. Applying the Lindeberg-Levy CLT to (8) shows $\sqrt{n}(\hat{\tau}-\tau) \xrightarrow{d} \mathcal{N}(0, \mathcal{V})$. ∎

The proof of Lemma 3 Recall the definition of $W(z)$. By Assumption 1(ii), it is true that $E[W(z)|Z, X] = E[W(z)|X]$, $z = 0, 1$. That is, if we treat Z as the treatment assignment and $W(z)$ as the potential outcomes, $W(z)$ and Z are unconfounded given X. Also, it is straightforward to check that Assumptions 1-5 of Thm. 1 of HIR are satisfied. The result for $\widehat{\Delta}$ follows directly from it. A similar argument applies to $\widehat{\Gamma}$. ∎

B. The proof of Theorem 3

Put $X = X_1 = X_2$. Under the conditions of Theorem 3, including one-sided non-compliance, the following hold:

$$\begin{aligned}
p =& P(D=1) = P(Z=1, D(1)=1) = P(D(1)=1|Z=1)P(Z=1) = \Gamma_t E(Z)\\
p(x) =& P(D=1|X=x) = P(D(1)=1|Z=1, X=x)P(Z=1|X=x) = \mu_1(x)q(x),\\
\mu_0(x) =& E[D|Z=0, X=x] = E[D(0)|X=x] = 0,\\
\rho_d(x) =& E[Y|D=d, X=x] = E[Y(d)|X=x],\ d=0,1.
\end{aligned}$$

Furthermore,

$$\begin{aligned}
m_1(x) =& E[Y|Z=1,X=x] \\
=& E[Y|Z=1,D=1,X]P[D=1|Z=1,X=x] \\
&+ E[Y|Z=1,D=0,X]P[D=0|Z=1,X=x] \\
=& E[Y(1)|Z=1,D=1,X=x]P[D(1)=1|Z=1,X=x] \\
&+ E[Y(0)|Z=1,D=0,X=x]P[D(1)=0|Z=1,X=x] \\
=& E[Y(1)|X=x]\mu_1(x) + E[Y(0)|X=x](1-\mu_1(x)) \\
=& \rho_1(x) - (1-\mu_1(x))(\rho_1(x)-\rho_0(x)),
\end{aligned}$$

where the fourth equality holds because the IV assumption and the unconfoundedness assumption hold jointly as in (12). Also,

$$\begin{aligned}
m_0(x) =& E[Y|Z=0,X=x] = E[Y|Z=0,D=0,X=x] \\
=& E[Y(0)|Z=0,D=0,X=x] = E[Y(0)|X=x] = \rho_0(x),
\end{aligned}$$

where the second equality holds since $D=0$ when $Z=0$. Define

$$\begin{aligned}
\phi_t(Y,D,X) &= \frac{p(X)}{p}\left\{\frac{D(Y-\rho_1(X))}{p(X)} - \frac{(1-D)(Y-\rho_0(X))}{(1-p(X))} + \frac{D(\rho_1(X)-\rho_0(X)-\beta_t)}{p(X)}\right\} \\
&\equiv \frac{p(X)}{p}\left\{\phi_1 - \phi_2 + \phi_3\right\},
\end{aligned}$$

and rewrite $\psi_t(Y,D,Z,X)$ as

$$\begin{aligned}
&\psi_t(Y,D,Z,X) \\
=& \frac{q(X)}{E(Z)\Gamma_t}\left\{\frac{Z\big[Y-m_1(X)-\tau_t(D-\mu_1(X))\big]}{q(X)} - \frac{(1-Z)\big[Y-m_0(X)-\tau_t(D-\mu_0(X))\big]}{1-q(X)}\right. \\
&\qquad\left. + \frac{Z\big[m_1(X)-m_0(X)-\tau_t(\mu_1(X)-\mu_0(X))\big]}{q(X)}\right\}, \\
=& \frac{p(X)}{p}\frac{1}{\mu_1(X)}\left\{\frac{Z(Y-\rho_1(X))}{q(X)} + \frac{Z(1-\mu_1(X))(\rho_1(X)-\rho_0(X))}{q(X)} - \frac{Z\tau_t(D-\mu_1(X))}{q(X)}\right. \\
&\qquad\left. - \frac{(1-Z)(Y-\rho_0(X))}{1-q(X)} + \frac{Z(\rho_1(X)-\rho_0(X)-\beta_t)\mu_1(X)}{q(X)}\right\} \\
=& \frac{p(X)}{p}\left\{\frac{Z(Y-\rho_1(X))}{p(X)} + \frac{Z(1-\mu_1(X))(\rho_1(X)-\rho_0(X))}{p(X)} - \frac{Z\beta_t((D-1)+(1-\mu_1(X)))}{p(X)}\right. \\
&\qquad\left. - \frac{(1-Z)(Y-\rho_0(X))}{(1-q(X))\mu_1(X)} + \frac{Z(\rho_1(X)-\rho_0(X)-\beta_t)\mu_1(X)}{p(X)}\right\} \\
=& \frac{p(X)}{p}\left\{\frac{Z(Y-\rho_1(X))}{p(X)} - \frac{(1-Z)(Y-\rho_0(X))}{(1-q(X))\mu_1(X)} + \frac{Z(\rho_1(X)-\rho_0(X)-\beta_t)}{p(X)} - \frac{Z\beta_t(D-1))}{p(X)}\right\} \\
\equiv& \frac{p(X)}{p}\left\{\psi_1 - \psi_2 + \psi_3 - \psi_4\right\}.
\end{aligned}$$

Note that

$$E[\phi_1\psi_1|X] = \frac{E\left[ZD(Y-\rho_1(X))^2\middle|X\right]}{p^2(X)} = \frac{E\left[D(Y-\rho_1(X))^2\middle|X\right]}{p^2(X)}$$
$$= \frac{E\left[D(Y-\rho_1(X))^2\middle|X, D=1\right]p(D=1|X=x)}{p^2(X)} = \frac{\sigma_1^2(X)}{p(X)},$$

where $\sigma_1^2(X) = V(Y(1)|X)$. Also, $E[\phi_1\psi_2] = 0$ since $(1-Z)D = 0$ with probability one and $E[\phi_1\psi_4|X] = 0$ since $D(1-D) = 0$. Note that

$$E[\phi_1\psi_3|X] = \frac{\rho_1(X)-\rho_0(X)-\beta_t}{p^2(X)}E[DZ(Y-\rho_1(X))|X]$$
$$= \frac{\rho_1(X)-\rho_0(X)-\beta_t}{p^2(X)}E[D(Y(1)-\rho_1(X))|X] = 0,$$

where the first equality in second line holds since $ZD = 1$ with probability one and the second equality holds since $E[D(Y(1)-\rho_1(X))|X] = 0$. Furthermore,

$$E[\phi_1\psi_1|X] = \frac{E[Z(1-D)(Y-\rho_1(X))(Y-\rho_0(X))|X]}{p(X)(1-p(X))}$$
$$= \frac{E[Z(1-D)(Y-\rho_1(X))(Y-\rho_0(X))|X, Z=1, D=0]P(Z=1, D=0|X)}{p(X)(1-p(X))}$$
$$= \frac{E[(Y(0)-\rho_1(X))(Y(0)-\rho_0(X))|X, Z=1, D=0](1-\mu_1(X))q(X)}{p(X)(1-p(X))}$$
$$= \frac{\sigma_0^2(X)(1-\mu_1(X))q(X)}{p(X)(1-p(X))} = \frac{\sigma_0^2(X)}{1-p(X)}\frac{1-\mu_1(X)}{\mu_1(X)},$$
$$E[\phi_1\psi_2|X] = \frac{E[(1-Z)(1-D)(Y-\rho_0(X))^2|X]}{(1-p(X))(1-q(X))\mu_1(X)}$$
$$= \frac{E[(1-Z)(1-D)(Y-\rho_0(X))^2|X, Z=0, D=0]P(Z=0, D=0|X)}{(1-p(X))(1-q(X))\mu_1(X)}$$
$$= \frac{E[(Y(0)-\rho_0(X))^2|X, D=0]P(Z=0|X)}{(1-p(X))(1-q(X))\mu_1(X)}$$
$$= \frac{\sigma_0^2(X)(1-q(X))}{(1-p(X))(1-q(X))\mu_1(X)} = \frac{\sigma_0^2(X)}{1-p(X)}\frac{1}{\mu_1(X)}.$$

Also, we have $E[\phi_2\psi_3] = 0$, $E[\phi_2\psi_4] = 0$, $E[\phi_3\psi_1] = 0$, $E[\phi_3\psi_2] = 0$ and $E[\phi_3\psi_4] = 0$. Finally,

$$E[\phi_3\psi_3] = \frac{E[D(\rho_1(X)-\rho_0(X)-\beta_t)^2|X]}{p^2(X)} = \frac{(\rho_1(X)-\rho_0(X)-\beta_t)^2}{p(X)}.$$

Consequently,

$$Cov(\phi_t, \psi_t) = E[\phi_t\psi_t]$$
$$= E\left[\frac{p^2(X)}{p}\left\{\frac{\sigma_1^2(X)}{p(X)} - \frac{\sigma_0^2(X)}{1-p(X)}\frac{1-\mu_1(X)}{\mu_1(X)} + \frac{\sigma_0^2(X)}{1-p(X)}\frac{1}{\mu_1(X)} + \frac{(\rho_1(X)-\rho_0(X)-\beta_t)^2}{p(X)}\right\}\right]$$
$$= E[\phi_t^2] = Var(\phi_t).$$

This shows Theorem 3. ∎

References

Abadie, A. (2003). Semiparametric Instrumental Variable Estimation of Treatment Response Models. *Journal of Econometrics 113*, 231–263.

Abadie, A., J. Angrist, and G. Imbens (2002). Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings. *Econometrica 70*, 91–117.

Abrevaya, J., Y.-C. Hsu, and R. P. Lieli (2012). Estimating Conditional Average Treatment Effects. Working paper.

Bloom, H. S. (1984). Accounting for No-Shows in Experimental Evaluation Designs. *Evaluation Review 8*, 225–246.

Bloom, H. S., L. L. Orr, S. H. Bell, G. Cave, F. Doolittle, W. Lin and J. M. Bos (1997). The Benefits and Costs of JTPA Title II-A Programs: Key Findings from the National Job Training Partnership Act Study. *Journal of Human Resources 32*, 549–576.

Chen, X., H. Hong, and A. Tarozzi (2008). Semiparametric Efficiency in GMM Models with Auxiliary Data. *Annals of Statistics 36*, 808–843.

Deaton, A. (2009). Instruments of Development: Randomization in the Tropics, and the Search for the Elusive Keys to Economic Development. *Proceedings of the British Academy, 2008 Lectures 162*, 123–160.

Donald, S. G. and Y.-C. Hsu (2012). Estimation and Inference for Distribution Functions and Quantile Functions in Treatment Effect Models. Working paper.

Frölich, M. (2007). Nonparametric IV Estimation of Local Average Treatment Effects with Covariates. *Journal of Econometrics 139*, 35–75.

Frölich, M. and M. Lechner (2010). Exploiting Regional Treatment Intensity for the Evaluation of Labour Market Policies. *Journal of American Statistical Association 105*, 1014–1029.

Frölich, M. and B. Melly (2008a). Identification of Treatment Effects on the Treated with One-Sided Non-Compliance. IZA Discussion Paper No. 3671.

Frölich, M. and B. Melly (2008b). Unconditional Quantile Treatment Effects under Endogeneity. IZA Discussion Paper No. 3288.

Guggenberger, P. (2010). The Impact of a Hausman Pretest on the Size of Hypothesis Tests. *Econometric Theory 26*, 369–382.

Hahn, J. (1998). On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects. *Econometrica 66*, 315–331.

Heckman, J. (1997). Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations. *The Journal of Human Resources 32*, 441–426.

Heckman, J., N. Hohmann, J. Smith, and M. Khoo (2000). Substitution and Dropout Bias in Social Experiments: A Study of an Influential Social Experiment. *The Quarterly Journal of Economics 115*, 651–694.

Heckman, J. and S. Urzúa (2010). Comparing IV with Structural Models: What Simple IV Can and Cannot Identify. *Journal of Econometrics 156*, 123–160.

Hirano, K., G. Imbens, and G. Ridder (2003). Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score. *Econometrica 71*, 1161–1189.

Hong, H. and D. Nekipelov (2010). Semiparametric Efficiency in Nonlinear LATE Models. *Quantitative Economics 1*, 279–304.

Ichimura, H. and O. Linton (2005). Asymptotic Expansions for Some Semiparametric Program Evaluation Estimators. In D. Andrews and J. Stock (Eds.), *Volume in Honor of Tom Rothenberg*. Cambridge University Press.

Imbens, G. (2010). Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzúa (2009). *Journal of Economic Literature 48*, 399–423.

Imbens, G. and J. Angrist (1994). Identification and Estimation of Local Average Treatment Effects. *Econometrica 62*, 467–475.

Imbens, G. W., W. Newey, and G. Ridder (2007). Mean-squared-error Calculations for Average Treatment Effects. Working paper.

Li, Q., J. S. Racine, and J. M. Wooldridge (2009). Efficient Estimation of Average Treatment Effects with Mixed Categorical and Continuous Data. *Journal of Business and Economic Statistics 27*, 206–223.

Wald, A. (1940). The Fitting of Straight Lines if Both Variables are Subject to Error. *Annals of Mathematical Statistics 11*, 284–300.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. The MIT Press; second edition.

Table 2: The distribution of $T = \sqrt{n}(\hat{\tau} - \tau)$ and $S = \sqrt{n}(\hat{\tau} - E\hat{\tau})/\widehat{s.e.}(\hat{\tau})$: Monte Carlo simulations

q	Series ($\hat{q}$)	Trim.	$n = 250$					$n = 500$					$n = 2500$				
			$\frac{E(T)}{\sqrt{n}}$	$sd(T)$	$E(\widehat{s.e.}(T))$	Prob. S < -1.645	Prob. S > 1.645	$\frac{E(T)}{\sqrt{n}}$	$s.e.(T)$	$E(\widehat{s.e.}(T))$	Prob. S < -1.645	Prob. S > 1.645	$\frac{E(T)}{\sqrt{n}}$	$s.e.(T)$	$E(\widehat{s.e.}(T))$	Prob. S < -1.645	Prob. S > 1.645
const.	const.	0.5%	-0.00	3.23	3.22	0.042	0.056	0.00	3.19	3.21	0.045	0.054	-0.00	3.21	3.19	0.049	0.052
	lin.	0.5%	-0.00	2.46	2.42	0.044	0.060	-0.00	2.45	2.40	0.044	0.059	-0.00	2.42	2.40	0.048	0.053
	quad.	0.5%	0.00	3.16	3.76	0.033	0.075	-0.00	2.58	2.51	0.048	0.067	-0.00	2.38	2.38	0.050	0.052
lin. 1	const.	0.5%	0.57	2.98	2.95	0.056	0.043	0.58	2.91	2.93	0.052	0.045	0.58	2.90	2.92	0.052	0.048
	lin.	0.5%	-0.01	3.54	2.72	0.074	0.110	-0.01	3.31	2.62	0.078	0.101	-0.00	3.11	2.59	0.082	0.089
	quad.	0.5%	-0.00	4.65	6.66	0.038	0.091	0.00	3.44	3.56	0.043	0.089	0.00	2.72	2.58	0.056	0.069
	cube	0.5%	0.00	12.4	85.7	0.002	0.027	0.01	6.17	20.10	0.015	0.043	0.00	2.93	2.94	0.051	0.078
lin. 2	const.	0.5%	0.45	4.07	3.99	0.063	0.017	0.45	3.82	3.82	0.063	0.025	0.44	3.72	3.73	0.058	0.041
	lin.	0.5%	-0.07	6.60	6.69	0.068	0.072	-0.03	5.57	4.77	0.064	0.099	-0.00	4.81	4.30	0.060	0.087
	lin.	5%	-0.29	7.32	4.93	0.065	0.180	-0.16	5.78	4.32	0.065	0.151	-0.07	4.91	4.09	0.065	0.099
	quad.	0.5%	-0.44	15.9	55.3	0.068	0.007	-0.08	8.40	21.6	0.068	0.009	-0.01	4.86	4.75	0.063	0.064
	quad.	5%	-1.17	16.0	11.9	0.074	0.161	-0.47	8.94	6.12	0.091	0.164	-0.12	5.57	4.09	0.092	0.135
	cube	0.5%	-5.42	180.9	788.0	0.088	0.040	-0.52	19.55	133.9	0.037	0.000	-0.02	6.22	15.57	0.043	0.006
	cube	5%	-8.19	159.7	73.4	0.097	0.318	-1.52	17.63	17.9	0.046	0.081	-0.25	6.55	4.74	0.101	0.142
Rat. 1	const.	0.5%	0.52	2.95	2.94	0.055	0.045	0.52	2.96	2.93	0.054	0.048	0.52	2.90	2.91	0.052	0.047
	lin.	0.5%	-0.12	3.29	2.73	0.055	0.107	-0.11	3.09	2.64	0.063	0.096	-0.11	2.97	2.59	0.070	0.081
	quad.	0.5%	-0.04	4.10	5.01	0.058	0.070	-0.04	3.20	2.87	0.065	0.079	-0.03	2.75	2.57	0.057	0.068
	quad.	5%	-0.05	3.93	4.1	0.056	0.076	-0.03	3.14	2.77	0.066	0.082	-0.03	2.76	2.57	0.055	0.069
	cube	0.5%	-0.08	11.2	70.9	0.003	0.015	-0.03	5.50	15.7	0.027	0.028	-0.01	3.11	3.26	0.068	0.050
	cube	5%	-0.09	7.10	18.6	0.007	0.016	-0.05	4.43	7.59	0.027	0.031	-0.02	2.99	2.71	0.067	0.073
Rat. 2	const.	0.5%	-0.31	3.21	3.21	0.034	0.064	-0.30	3.21	3.20	0.039	0.058	-0.30	3.19	3.18	0.047	0.055
	lin.	0.5%	0.01	2.46	2.39	0.044	0.060	0.00	2.42	2.36	0.048	0.058	0.01	2.34	2.34	0.047	0.057
	quad.	0.5%	-0.00	3.02	3.36	0.042	0.061	0.00	2.46	2.40	0.047	0.061	0.00	2.35	2.31	0.052	0.055
	cube	0.5%	-0.03	8.64	53.54	0.004	0.014	0.00	4.12	11.13	0.022	0.030	-0.00	2.39	2.35	0.054	0.055

Table 3: Properties of the unconfoundedness test and the distribution of $\sqrt{n}[(\widehat{L)ATT} - (L)ATT]$ for various estimators: $b = 0$ (unconfoundedness holds)

				$n = 250$				$n = 500$				$n = 2500$			
q	Series ($\hat{q}$)	Trim.	Estimator	$\frac{\text{Mean}}{\sqrt{n}}$	s.e.	E($\widehat{s.e.}$)	MSE	$\frac{\text{Mean}}{\sqrt{n}}$	s.e.	E($\widehat{s.e.}$)	MSE	$\frac{\text{Mean}}{\sqrt{n}}$	s.e.	E($\widehat{s.e.}$)	MSE
Const.	lin.	.5%	$\hat{\tau}_t$ (LATT)	-0.00	2.64	2.54	6.96	0.00	2.47	2.47	6.11	-0.00	2.42	2.42	5.88
			$\hat{\beta}_t$ (ATT)	-0.02	1.63	1.46	2.73	-0.01	1.56	1.44	2.51	-0.01	1.54	1.44	2.91
			Combined	-0.01	1.62	1.45	2.66	-0.01	1.56	1.44	2.49	-0.01	1.53	1.44	2.89
			Pre-tested	-0.01	1.99	1.52	3.97	-0.01	1.88	1.50	3.55	-0.01	1.92	1.51	3.96
			Size/$E(\hat{a})$			0.061/0.02				0.057/0.01				0.070/0.01	
Const.	quad.	.5%	$\hat{\tau}_t$ (LATT)	0.01	3.75	5.04	14.07	0.00	2.76	2.78	7.63	0.00	2.48	2.45	6.15
			$\hat{\beta}_t$ (ATT)	0.01	2.04	1.88	4.17	-0.00	1.61	1.50	2.60	0.00	1.46	1.44	2.12
			Combined	0.02	1.93	1.67	3.77	-0.00	1.59	1.49	2.54	0.00	1.46	1.44	2.12
			Pre-tested	0.03	2.34	1.74	5.63	0.00	1.94	1.54	3.78	-0.00	1.79	1.49	3.22
			Size/$E(\hat{a})$			0.045/0.07				0.046/0.03				0.048/0.01	
Lin. 1	lin.	.5%	$\hat{\tau}_t$ (LATT)	0.00	4.42	3.04	19.52	0.00	4.11	2.78	16.86	0.00	3.91	2.63	15.30
			$\hat{\beta}_t$ (ATT)	-0.01	1.87	1.53	3.50	-0.01	1.78	1.51	3.18	-0.01	1.71	1.51	3.01
			Combined	0.01	1.84	1.51	3.40	0.00	1.77	1.51	3.14	-0.00	1.71	1.51	2.97
			Pre-tested	0.02	3.24	1.75	10.63	0.00	3.43	1.77	11.76	0.00	3.36	1.76	11.28
			Size/$E(\hat{a})$			0.157/0.03				0.192/0.02				0.220/0.01	
Lin. 1	quad.	.5%	$\hat{\tau}_t$ (LATT)	0.02	6.70	9.62	45.03	0.01	4.27	4.24	18.30	0.00	3.02	2.73	9.14
			$\hat{\beta}_t$ (ATT)	0.01	2.52	2.19	6.36	0.00	1.86	1.59	3.49	0.00	1.54	1.50	2.38
			Combined	0.03	2.29	1.80	5.45	0.01	1.83	1.56	3.37	0.00	1.54	1.50	2.38
			Pre-tested	0.07	3.48	1.98	13.3	0.02	2.67	1.68	7.43	0.00	2.23	1.61	4.98
			Size/$E(\hat{a})$			0.092/0.06				0.085/0.03				0.079/0.01	
Lin. 1	cube	.5%	$\hat{\tau}_t$ (LATT)	0.21	18.04	112.42	336.26	0.04	10.41	40.22	109.01	0.00	3.49	3.66	12.19
			$\hat{\beta}_t$ (ATT)	0.08	11.88	48.16	142.77	0.03	3.76	5.99	14.44	0.00	1.65	1.56	2.83
			Combined	0.29	5.98	8.04	57.09	0.06	3.02	2.82	11.05	0.00	1.67	1.55	2.81
			Pre-tested	0.36	7.42	8.26	87.20	0.09	4.14	2.97	21.37	0.01	2.26	1.64	5.21
			Size/$E(\hat{a})$			0.081/0.18				0.072/0.11				0.061/0.02	
Lin. 2	lin.	.5%	$\hat{\tau}_t$ (LATT)	0.00	4.46	3.81	19.87	0.00	4.09	3.68	16.69	0.00	3.87	3.60	14.96
			$\hat{\beta}_t$ (ATT)	-0.00	2.78	2.47	7.72	-0.00	2.64	2.49	6.99	-0.00	2.57	2.50	6.59
			Combined	0.00	2.72	2.46	7.43	0.00	2.63	2.49	6.90	-0.00	2.56	2.50	6.58
			Pre-tested	0.00	3.39	2.57	11.52	0.00	3.18	2.57	10.13	-0.00	3.11	2.58	9.68
			Size/$E(\hat{a})$			0.063/0.01				0.064/0.01				0.072/0.00	
Lin. 2	quad.	.5%	$\hat{\tau}_t$ (LATT)	0.01	6.49	4.84	42.18	0.01	4.47	3.87	20.03	0.00	3.81	3.63	14.53
			$\hat{\beta}_t$ (ATT)	0.01	12.10	4.96	146.54	0.01	2.89	2.53	8.39	0.00	2.55	2.50	6.49
			Combined	0.05	3.97	2.70	16.41	0.01	2.82	2.51	8.03	0.00	2.55	2.50	6.50
			Pre-tested	0.04	4.76	2.92	23.13	0.01	3.48	2.63	12.16	0.00	3.01	2.58	9.06
			Size/$E(\hat{a})$			0.120/0.10				0.080/0.02				0.067/0.01	
Lin. 2	cube	.5%	$\hat{\tau}_t$ (LATT)	0.22	81.70	293.30	6686.90	0.06	10.35	10.98	109.14	0.00	4.07	3.83	16.60
			$\hat{\beta}_t$ (ATT)	2.24	38.69	85.22	2753.10	0.10	22.82	25.39	526.13	0.00	2.64	2.54	6.98
			Combined	1.81	11.34	6.42	948.94	0.18	5.68	3.23	48.71	0.00	2.62	2.54	6.92
			Pre-tested	1.65	25.13	9.06	1314.86	0.15	6.20	3.54	50.36	0.00	3.14	2.62	9.89
			Size/$E(\hat{a})$			0.606/0.27				0.163/0.18				0.070/0.01	

Table 4: Properties of the unconfoundedness test and the distribution of $\sqrt{n}[\widehat{(L)ATT} - (L)ATT]$ for various estimators: $b = 0$ (unconfoundedness holds)

				$n = 250$				$n = 500$				$n = 2500$			
q	Series ($\hat{q}$)	Trim.	Estimator	$\frac{\text{Mean}}{\sqrt{n}}$	s.e.	E($\widehat{s.e.}$)	MSE	$\frac{\text{Mean}}{\sqrt{n}}$	s.e.	E($\widehat{s.e.}$)	MSE	$\frac{\text{Mean}}{\sqrt{n}}$	s.e.	E($\widehat{s.e.}$)	MSE
Lin. 2	lin.	5%	$\hat{\tau}_t$ (LATT)	-0.08	5.05	3.85	27.04	-0.04	4.50	3.69	20.96	-0.02	4.07	3.60	17.14
			$\hat{\beta}_t$ (ATT)	0.11	1.97	2.06	6.89	0.11	1.93	2.14	9.32	0.10	1.90	2.20	29.00
			Combined	0.11	2.00	2.05	7.26	0.11	1.92	2.13	10.10	0.11	1.89	2.19	33.40
			Pre-tested	0.01	4.69	2.46	22.04	0.03	4.49	2.55	20.48	0.01	5.16	3.01	26.92
			Size/$E(\hat{a})$		0.192/-0.06				0.243/-0.07				0.570/-0.07		
Lin. 2	quad.	5%	$\hat{\tau}_t$ (LATT)	-0.29	7.34	4.95	75.41	-0.13	5.20	3.92	35.16	-0.03	4.07	3.63	18.45
			$\hat{\beta}_t$ (ATT)	0.12	3.86	2.35	18.51	0.11	2.18	2.06	11.10	0.10	1.89	2.17	30.25
			Combined	0.12	3.44	2.05	15.67	0.12	2.17	2.04	12.24	0.11	1.87	2.16	35.41
			Pre-tested	-0.16	7.16	3.20	57.43	-0.06	5.80	2.89	35.15	-0.00	5.19	3.12	26.96
			Size/$E(\hat{a})$		0.386/0.01				0.420/-0.03				0.645/-0.08		
Lin. 2	cube	5%	$\hat{\tau}_t$ (LATT)	-0.05	25.12	27.45	631.63	-0.33	9.28	8.40	141.13	-0.07	4.47	3.72	31.01
			$\hat{\beta}_t$ (ATT)	1.59	6.82	3.35	679.48	0.19	5.39	3.90	47.83	0.11	2.04	2.15	34.34
			Combined	1.44	6.31	2.60	561.64	0.18	5.51	2.48	46.02	0.12	2.07	2.12	41.36
			Pre-tested	0.71	19.71	6.45	513.24	-0.17	8.63	4.15	89.43	-0.05	5.44	3.44	36.43
			Size/$E(\hat{a})$		0.624/0.08				0.545/0.09				0.818/-0.07		
Rat. 1	lin.	.5%	$\hat{\tau}_t$ (LATT)	-0.17	4.20	3.25	24.89	-0.17	4.00	3.03	30.84	-0.17	3.76	2.89	88.39
			$\hat{\beta}_t$ (ATT)	-0.02	1.86	1.54	3.58	-0.02	1.78	1.53	3.41	-0.02	1.73	1.53	4.39
			Combined	-0.01	1.85	1.53	3.46	-0.01	1.77	1.52	3.22	-0.01	1.74	1.52	3.49
			Pre-tested	-0.07	3.43	1.84	13.12	-0.11	4.15	2.08	23.59	-0.16	4.41	2.69	85.78
			Size/$E(\hat{a})$		0.151/-0.02				0.305/-0.04				0.833/-0.06		
Rat. 1	quad.	.5%	$\hat{\tau}_t$ (LATT)	0.01	4.37	4.84	19.13	-0.01	3.21	2.96	10.29	-0.01	2.63	2.51	7.26
			$\hat{\beta}_t$ (ATT)	0.01	2.26	1.91	5.14	-0.00	1.74	1.52	3.04	-0.00	1.51	1.46	2.28
			Combined	0.02	2.08	1.69	4.43	0.00	1.71	1.51	2.94	-0.00	1.51	1.46	2.29
			Pre-tested	0.04	2.77	1.80	8.07	0.00	2.32	1.61	5.39	-0.01	1.99	1.52	4.04
			Size/$E(\hat{a})$		0.071/0.07				0.073/0.03				0.064/0.01		
Rat. 1	cube	.5%	$\hat{\tau}_t$ (LATT)	0.15	16.11	91.49	265.28	0.02	8.42	27.82	71.21	-0.00	2.86	2.79	8.22
			$\hat{\beta}_t$ (ATT)	0.08	11.06	43.62	124.18	0.02	3.28	4.50	11.02	-0.00	1.60	1.51	2.56
			Combined	0.27	5.58	6.57	49.64	0.05	2.74	2.57	8.92	-0.00	1.58	1.50	2.52
			Pre-tested	0.31	6.39	6.72	65.22	0.07	3.37	2.66	13.77	-0.00	2.03	1.57	4.11
			Size/$E(\hat{a})$		0.058/0.19				0.052/0.12				0.060/0.02		
Rat. 2	lin.	.5%	$\hat{\tau}_t$ (LATT)	-0.01	2.56	2.49	6.60	-0.02	2.50	2.45	6.36	-0.01	2.46	2.43	6.52
			$\hat{\beta}_t$ (ATT)	-0.01	1.67	1.52	2.84	-0.02	1.59	1.51	2.65	-0.01	1.54	1.51	2.79
			Combined	-0.02	1.66	1.51	2.82	-0.02	1.59	1.51	2.65	-0.01	1.54	1.51	2.80
			Pre-tested	-0.01	1.93	1.57	3.78	-0.02	1.91	1.56	3.77	-0.01	1.87	1.56	3.94
			Size/$E(\hat{a})$		0.053/0.03				0.053/0.02				0.054/0.02		
Rat. 2	quad.	.5%	$\hat{\tau}_t$ (LATT)	0.00	3.37	4.14	11.36	-0.00	2.72	2.68	7.43	-0.00	2.47	2.49	6.13
			$\hat{\beta}_t$ (ATT)	0.01	1.94	1.79	3.78	0.00	1.67	1.56	2.78	-0.00	1.55	1.52	2.42
			Combined	0.01	1.88	1.68	3.57	0.00	1.66	1.55	2.77	-0.00	1.55	1.52	2.42
			Pre-tested	0.02	2.23	1.74	5.03	0.00	1.98	1.61	3.90	-0.00	1.85	1.57	3.44
			Size/$E(\hat{a})$		0.048/0.06				0.048/0.03				0.049/0.01		
Rat. 2	cube	.5%	$\hat{\tau}_t$ (LATT)	0.09	10.41	59.04	110.67	0.01	4.67	13.00	21.94	-0.00	2.63	2.55	6.95
			$\hat{\beta}_t$ (ATT)	0.08	6.54	21.97	44.38	0.01	2.57	3.69	6.68	-0.00	1.61	1.54	2.64
			Combined	0.19	3.98	4.60	24.99	0.03	2.27	2.22	5.61	-0.00	1.61	1.54	2.64
			Pre-tested	0.21	4.43	4.66	30.77	0.04	2.59	2.27	7.44	-0.00	1.98	1.60	3.94
			Size/$E(\hat{a})$		0.037/0.19				0.036/0.12				0.054/0.01		

Table 5: Properties of the unconfoundedness test and the distribution of $\sqrt{n}[\widehat{(L)ATT} - (L)ATT]$ for various estimators: $b = 0.5$ (unconfoundedness does not hold)

				$n = 250$				$n = 500$				$n = 2500$			
q	Series ($\hat{q}$)	Trim.	Estimator	$\frac{\text{Mean}}{\sqrt{n}}$	s.e.	E($\widehat{s.e.}$)	MSE	$\frac{\text{Mean}}{\sqrt{n}}$	s.e.	E($\widehat{s.e.}$)	MSE	$\frac{\text{Mean}}{\sqrt{n}}$	s.e.	E($\widehat{s.e.}$)	MSE
Const.	lin.	.5%	$\hat{\tau}_t$ (LATT)	-0.01	2.63	2.57	6.91	-0.00	2.51	2.50	6.31	0.00	2.47	2.46	6.11
			$\hat{\beta}_t$ (ATT)	0.26	2.16	1.64	21.40	0.26	2.03	1.60	37.45	0.26	1.95	1.58	173.48
			Combined	0.24	2.21	1.61	19.81	0.24	2.14	1.59	34.19	0.25	2.07	1.58	156.90
			Pre-tested	0.05	3.21	2.15	11.06	0.02	3.03	2.33	9.37	0.00	2.47	2.46	6.11
			Power/$E(\hat{a})$			0.509/0.09				0.785/0.07				1.000/0.05	
Const.	quad.	.5%	$\hat{\tau}_t$ (LATT)	0.00	3.84	5.04	14.72	-0.00	2.82	2.81	7.97	-0.00	2.48	2.48	6.16
			$\hat{\beta}_t$ (ATT)	0.35	2.36	2.16	36.66	0.35	1.78	1.61	63.93	0.35	1.53	1.51	300.43
			Combined	0.33	2.50	1.82	33.78	-0.33	1.97	1.58	59.68	0.34	1.60	1.51	296.95
			Pre-tested	0.18	4.05	2.38	24.13	0.01	3.26	2.68	10.70	-0.00	2.48	2.48	6.16
			Power/$E(\hat{a})$			0.344/0.10				0.912/0.05				1.000/0.01	
Lin. 1	quad.	.5%	$\hat{\tau}_t$ (LATT)	0.03	6.11	8.58	37.56	0.01	4.24	4.26	17.98	0.00	3.06	2.78	9.38
			$\hat{\beta}_t$ (ATT)	0.35	2.99	2.65	39.62	0.34	2.14	1.72	62.37	0.34	1.63	1.50	288.81
			Combined	0.35	2.78	1.97	38.45	0.33	2.28	1.63	60.77	0.34	1.73	1.49	298.08
			Pre-tested	0.28	4.01	2.31	35.11	0.10	4.85	2.74	28.54	0.00	3.06	2.78	9.38
			Power/$E(\hat{a})$			0.160/0.09				0.606/0.04				1.000/0.00	
Lin. 1	cube	.5%	$\hat{\tau}_t$ (LATT)	0.21	18.31	113.72	346.58	0.05	9.03	28.56	82.81	0.00	3.44	3.56	11.82
			$\hat{\beta}_t$ (ATT)	0.43	13.68	58.98	233.97	0.36	5.10	11.21	90.10	0.34	1.79	1.65	292.16
			Combined	0.61	7.23	9.19	146.40	0.36	4.08	3.45	81.85	0.33	2.52	1.60	286.83
			Pre-tested	0.62	7.45	9.23	152.31	0.33	4.69	3.64	77.97	0.01	3.88	3.35	15.15
			Power/$E(\hat{a})$			0.017/0.22				0.068/0.14				0.986/0.02	
Lin. 2.	lin.	.5%	$\hat{\tau}_t$ (LATT)	-0.02	4.59	3.95	21.13	-0.01	4.26	3.81	18.19	-0.00	3.94	3.72	15.53
			$\hat{\beta}_t$ (ATT)	0.25	2.94	2.47	24.35	0.25	2.66	2.48	38.71	0.25	2.55	2.49	164.11
			Combined	0.25	2.88	2.46	24.22	0.25	2.66	2.47	39.21	0.25	2.56	2.48	167.73
			Pre-tested	0.12	4.85	2.95	27.14	0.06	5.15	3.28	28.30	-0.00	3.96	3.72	15.71
			Power/$E(\hat{a})$			0.258/0.01				0.539/-0.00				0.998/-0.01	
Lin. 2	quad.	.5%	$\hat{\tau}_t$ (LATT)	0.01	6.27	4.84	39.38	-0.01	4.62	3.98	21.39	0.00	3.78	3.73	14.28
			$\hat{\beta}_t$ (ATT)	0.28	10.02	5.64	120.37	0.27	2.98	2.51	45.07	0.26	2.50	2.47	179.00
			Combined	0.29	4.16	2.72	38.19	0.27	2.96	2.49	44.33	0.27	2.51	2.47	183.05
			Pre-tested	0.14	5.61	3.29	36.38	0.06	5.38	3.37	30.90	0.00	3.80	3.73	14.44
			Power/$E(\hat{a})$			0.270/0.12				0.536/0.02				0.998/0.01	
Lin. 2	cube	.5%	$\hat{\tau}_t$ (LATT)	0.20	84.96	303.79	7227.8	0.06	9.08	9.61	84.10	-0.00	4.03	3.82	16.26
			$\hat{\beta}_t$ (ATT)	2.63	30.37	46.76	2656.1	0.37	19.90	18.49	464.75	0.27	2.69	2.50	183.19
			Combined	2.06	11.08	5.77	1188.3	0.40	6.24	3.28	117.43	0.27	2.77	2.49	184.29
			Pre-tested	1.66	25.07	8.71	1316.81	0.20	7.14	4.10	71.58	-0.00	4.05	3.81	16.43
			Power/$E(\hat{a})$			0.677/0.27				0.379/0.24				0.998/-0.01	
Rat. 1	quad.	.5%	$\hat{\tau}_t$ (LATT)	-0.01	4.52	5.31	20.47	-0.02	3.24	2.99	10.68	-0.01	2.67	2.56	7.67
			$\hat{\beta}_t$ (ATT)	0.33	2.68	2.20	33.89	0.32	1.90	1.59	54.11	0.32	1.57	1.48	255.50
			Combined	0.31	2.69	1.81	31.95	0.31	2.10	1.56	51.73	0.32	1.65	1.48	256.86
			Pre-tested	0.17	4.26	2.36	25.37	0.01	3.89	2.74	15.14	-0.01	2.67	2.56	7.67
			Power/$E(\hat{a})$			0.294/0.10				0.837/0.04				1.000/0.00	
Rat. 1	cube	.5%	$\hat{\tau}_t$ (LATT)	0.16	15.24	81.02	238.97	0.03	6.60	15.72	43.93	-0.00	2.90	2.86	8.43
			$\hat{\beta}_t$ (ATT)	0.41	13.97	57.61	236.90	0.35	3.82	5.93	76.42	0.32	1.68	1.56	264.69
			Combined	0.57	6.60	7.79	123.63	0.34	3.66	2.76	71.01	0.32	2.18	1.54	259.18
			Pre-tested	0.57	6.76	7.84	126.31	0.27	4.62	3.12	59.14	-0.00	3.04	2.80	9.24
			Power/$E(\hat{a})$			0.020/0.25				0.157/0.15				0.995/0.02	
Rat. 2	lin.	.5%	$\hat{\tau}_t$ (LATT)	-0.02	2.60	2.54	6.85	-0.02	2.57	2.50	6.74	-0.02	2.51	2.47	6.89
			$\hat{\beta}_t$ (ATT)	0.26	2.00	1.67	20.48	0.26	1.93	1.65	36.41	0.26	1.88	1.64	169.11
			Combined	0.24	2.03	1.65	18.32	0.24	1.98	1.64	32.26	0.24	1.93	1.64	148.81
			Pre-tested	0.04	3.21	2.20	10.66	-0.00	3.02	2.39	9.11	-0.02	2.51	2.47	6.89
			Power/$E(\hat{a})$			0.571/0.08				0.854/0.07				1.000/0.06	
Rat. 2	quad.	.5%	$\hat{\tau}_t$ (LATT)	0.00	3.36	3.88	11.27	-0.00	2.73	2.69	7.44	-0.00	2.58	2.53	6.70
			$\hat{\beta}_t$ (ATT)	0.34	2.20	2.04	33.16	0.33	1.82	1.68	58.57	0.33	1.68	1.62	272.24
			Combined	0.31	2.31	1.81	29.82	0.32	1.91	1.66	53.72	0.32	1.72	1.62	260.59
			Pre-tested	0.12	3.91	2.41	19.10	0.01	3.08	2.61	9.52	-0.00	2.58	2.53	6.70
			Power/$E(\hat{a})$			0.462/0.10				0.936/0.06				1.000/0.02	
Rat. 2	cube	.5%	$\hat{\tau}_t$ (LATT)	0.08	10.93	67.04	121.21	0.02	4.58	11.24	21.16	-0.00	2.64	2.61	7.02
			$\hat{\beta}_t$ (ATT)	0.40	9.12	32.80	123.83	0.35	2.90	4.34	70.17	0.33	1.69	1.66	273.89
			Combined	0.48	4.74	5.69	79.06	0.32	3.37	2.47	62.33	0.32	1.84	1.65	255.62
			Pre-tested	0.47	4.80	5.74	77.13	0.20	4.67	3.03	42.20	-0.00	2.65	2.61	7.06
			Power/$E(\hat{a})$			0.020/0.24				0.326/0.17				0.999/0.03	

Table 6: Properties of the unconfoundedness test and the distribution of $\sqrt{n}[\widehat{(L)ATT} - (L)ATT]$ for various estimators: $b = 0.25$ (unconfoundedness does not hold)

				$n = 250$				$n = 500$				$n = 2500$			
q	Series ($\hat{q}$)	Trim.	Estimator	$\frac{\text{Mean}}{\sqrt{n}}$	s.e.	E($\widehat{s.e.}$)	MSE	$\frac{\text{Mean}}{\sqrt{n}}$	s.e.	E($\widehat{s.e.}$)	MSE	$\frac{\text{Mean}}{\sqrt{n}}$	s.e.	E($\widehat{s.e.}$)	MSE
Const.	quad.	.5%	$\hat{\tau}_t$ (LATT)	0.00	3.85	5.14	14.84	-0.00	2.81	2.80	7.89	-0.00	2.49	2.45	6.20
			$\hat{\beta}_t$ (ATT)	0.16	2.18	2.09	10.88	0.15	1.71	1.56	14.68	0.15	1.50	1.48	60.92
			Combined	0.15	2.10	1.76	10.38	0.15	1.74	1.54	14.07	0.15	1.51	1.48	60.38
			Pre-tested	0.13	2.64	1.88	11.06	0.07	3.17	1.92	12.43	0.00	2.67	2.43	7.16
			Power/$E(\hat{a})$			0.075/0.08				0.307/0.04				0.974/0.00	
Lin. 1	quad.	.5%	$\hat{\tau}_t$ (LATT)	0.02	7.32	11.41	53.69	0.01	4.42	4.40	19.53	-0.00	3.06	2.75	9.37
			$\hat{\beta}_t$ (ATT)	0.16	2.72	2.40	13.64	0.15	2.04	1.67	15.53	0.15	1.60	1.51	56.50
			Combined	0.17	2.56	1.94	14.00	0.15	1.99	1.61	15.42	0.15	1.62	1.51	57.21
			Pre-tested	0.17	3.24	2.06	17.98	0.11	3.28	1.89	16.41	0.01	3.67	2.60	13.57
			Power/$E(\hat{a})$			0.062/0.08				0.158/0.04				0.874/0.00	
Lin. 2	quad.	.5%	$\hat{\tau}_t$ (LATT)	0.00	6.71	5.19	44.97	0.01	4.50	3.88	20.28	0.00	3.80	3.65	14.48
			$\hat{\beta}_t$ (ATT)	0.13	9.33	4.95	91.47	0.13	2.93	2.53	16.63	0.12	2.58	2.50	40.75
			Combined	0.16	4.14	2.70	23.21	0.13	2.87	2.51	16.44	0.12	2.59	2.50	41.09
			Pre-tested	0.09	5.03	3.01	27.50	0.07	4.12	2.79	19.57	0.03	4.76	3.18	24.92
			Power/$E(\hat{a})$			0.154/0.11				0.176/0.02				0.578/0.00	
Rat. 1	quad.	.5%	$\hat{\tau}_t$ (LATT)	0.00	4.66	5.48	21.69	-0.01	3.24	2.96	10.58	-0.01	2.62	2.52	7.32
			$\hat{\beta}_t$ (ATT)	0.15	2.52	2.08	11.92	0.14	1.86	1.58	13.13	0.14	1.55	1.48	49.87
			Combined	0.15	2.34	1.78	11.33	0.14	1.88	1.55	12.79	0.14	1.58	1.48	49.29
			Pre-tested	0.13	3.00	1.93	12.99	0.06	3.41	1.95	13.20	-0.01	2.91	2.48	8.72
			Power/$E(\hat{a})$			0.085/0.08				0.279/0.04				0.955	
Rat. 2	quad.	.5%	$\hat{\tau}_t$ (LATT)	-0.00	3.32	3.83	11.00	-0.00	2.69	2.67	7.23	-0.00	2.50	2.49	6.29
			$\hat{\beta}_t$ (ATT)	0.15	2.07	1.89	9.68	0.14	1.74	1.62	13.46	0.14	1.60	1.57	53.72
			Combined	0.14	2.01	1.74	9.11	0.14	1.75	1.62	12.72	0.14	1.61	1.57	52.42
			Pre-tested	0.11	2.65	1.88	9.83	0.06	3.12	1.96	11.53	-0.00	2.74	2.46	7.50
			Power/$E(\hat{a})$			0.110/0.77				0.328/0.04				0.964/0.01	
Rat. 2	cube	.5%	$\hat{\tau}_t$ (LATT)	0.10	10.19	54.80	106.31	0.02	4.33	9.80	18.95	0.00	2.68	2.57	7.18
			$\hat{\beta}_t$ (ATT)	0.22	8.04	30.86	76.68	0.17	2.71	4.11	21.02	0.14	1.65	1.62	54.22
			Combined	0.32	4.50	5.46	46.16	0.16	2.56	2.33	19.99	0.14	1.70	1.60	51.63
			Pre-tested	0.33	4.64	5.48	48.20	0.15	2.92	2.43	19.35	0.01	3.05	2.50	9.39
			Power/$E(\hat{a})$			0.014/0.23				0.061/0.15				0.931/0.01	