Review

Martin Huber[1] / Kaspar Wüthrich[2]

Local Average and Quantile Treatment Effects Under Endogeneity: A Review

[1] University of Fribourg, Bd. de Pérolles 90, 1700 Fribourg, Switzerland, E-mail: martin.huber@unifr.ch
[2] UC San Diego, Department of Economics, San Diego, 9500 Gilman Dr. La Jolla, CA92093, USA

Abstract:
This paper provides a review of methodological advancements in the evaluation of heterogeneous treatment effect models based on instrumental variable (IV) methods. We focus on models that achieve identification by assuming monotonicity of the treatment in the IV and analyze local average and quantile treatment effects for the subpopulation of compliers. We start with a comprehensive discussion of the binary treatment and binary IV case as for instance relevant in randomized experiments with imperfect compliance. We then review extensions to identification and estimation with covariates, multi-valued and multiple treatments and instruments, outcome attrition and measurement error, and the identification of direct and indirect treatment effects, among others. We also discuss testable implications and possible relaxations of the IV assumptions, approaches to extrapolate from local to global treatment effects, and the relationship to other IV approaches.

Keywords: instrument, LATE, selection on unobservables, treatment effects

JEL classification: C26

DOI: 10.1515/jem-2017-0007

1 Introduction

In empirical research, the assessment of the causal effect of a treatment (e.g. training or education) on an outcome (e.g. earnings) is frequently complicated by endogeneity, implying that the treatment is not as good as randomly assigned. That is, individuals may select themselves into the treatment in a non-random way that is related to their expected gains from the treatment in the outcome. This happens e.g. in experiments with non-compliance in which access to the treatment is randomly assigned, but some individuals do not comply with the randomization and choose a different treatment state. If compliance behaviour is associated with unobserved characteristics (e.g. ability) that also affect the outcome, endogeneity jeopardizes a causal analysis based on simple comparisons between treated and non-treated observations.

Causal effects can nevertheless be identified in the presence of an instrumental variable (IV) that (i) affects the treatment decision of (at least) some subjects and (ii) is otherwise not associated with the potential outcomes under either treatment state. For this reason, IV methods have become a cornerstone of causal inference. They represent an alternative strategy to identification based on selection on observables, see for instance the surveys in Imbens (2004) and Imbens and Wooldridge (2009), where the treatment decision is assumed to be as good as random after controlling for observed characteristics. While the latter assumption permits identifying the average treatment effect in the total population (ATE) or on the treated (ATT), it seems implausibly restrictive in many applications, as subjects might select themselves into treatment based on unobserved information about their potential outcomes and treatment gains.

This paper reviews the methodological advancements in the IV-based evaluation of treatment effects under effect heterogeneity. This implies that treatment effectiveness may vary across subjects as a function of observed and unobserved characteristics. In such models, two stage least squares (TSLS) consistently estimates the average treatment effect for the compliant subpopulation, the so-called local average treatment effect (LATE), given that the treatment and instrument are binary and the treatment is weakly monotonic in the instrument. In experiments, compliers are those whose treatment status is induced by the assignment. That is, they take the treatment when randomized in, but abstain from it when randomized out.

Following the seminal paper of Imbens and Angrist (1994), much progress has been made in extending the initial framework in various empirically relevant dimensions. This includes the evaluation of distributional and quantile treatment effects, multivalued or multiple treatments and instruments, identification and estimation

Martin Huber is the corresponding author.

in the presence of observed covariates, attrition and measurement error, and more. Furthermore, it has been acknowledged that the LATE assumptions have testable implications that may be verified in the data and that causal effects might be point or partially identified under weaker conditions. Finally, conditions and tests for the external validity of the LATE with respect to the ATE have been proposed. This appears important in the light of the controversial debate about the relevance of the complier population, see for instance the discussions in Deaton (2010), Imbens (2010), and Heckman and Urzúa (2010).

Our review complements more introductory surveys of the LATE framework, see Imbens (2014) and the textbook discussions in Angrist and Pischke (2009, 2015). A more specialized review focussing on the specific aspects of identifying and estimating the local quantile treatment effect (LQTE) is provided by Melly and Wüthrich (2018).

The survey is structured as follows. Section 2 reviews the IV assumptions in the binary instrument and treatment case and the identification of the LATE, LQTE, and potential outcome means and distributions. It also discusses identification under multivalued treatments and instruments and considers the concept of marginal treatment effects. Section 3 considers a conditional version of the IV assumptions in the presence of covariates along with the identification of local, quantile, and marginal treatment effects as well as more general functionals among compliers. Section 4 discusses extensions of the IV framework to more complex identification problems, including non-response bias in the outcome, measurement error in the treatment or the instrument, the presence of dynamic, i.e. sequentially assigned, or multiple treatments, and the evaluation of causal mechanisms (or direct and indirect effects) of the treatment. Section 5 discusses how violations of the IV assumptions affect identification and under which relaxations of the assumptions causal effects on specific subpopulations can nevertheless be obtained. Section 6 outlines approaches for testing the IV assumptions and briefly discusses sensitivity checks and bounds analysis under specific violations of the assumptions. Section 7 is concerned with the external validity of the LATE for the entire population. It discusses potential checks for external validity based on observables, conditions for extrapolating the LATE to the ATE along with testable implications, and partial identification of the ATE based on the IV assumptions and possibly further restrictions. Section 8 clarifies the relationship of the framework considered in this paper and other IV approaches suggested in the literature. Specifically, we discuss the connection to the classical linear IV model with covariates and to the instrumental variable quantile regression model (Chernozhukov and Hansen 2005). Section 9 concludes.

2 Identification and Estimation without Covariates

We first consider a setup with a binary treatment and instrument. Section 2.1 discusses the IV assumptions, while Section 2.2 shows the identification of the LATE, LQTE, and the potential outcome distributions among compliers. Section 2.4 extends the setup to multivalued treatments. Section 2.3 considers multivalued IVs as well as the marginal treatment effect.

2.1 Assumptions

Treatment evaluation typically considers the effect of a binary treatment $D \in \{1, 0\}$. Examples include receiving or not receiving a labor market policy (e.g. training) or a medical treatment. Y denotes the outcome on which the effect ought to be estimated, for instance, employment or earnings, measured at some point in time after the treatment. Under endogeneity, unobserved factors affect both D and Y such that the treatment effect is not obtained from simply comparing treatment and the control group. However, if there exists an instrument Z which is relevant (i.e. influences the treatment status) and valid (i.e. not associated with the unobserved factors) and does not directly affect the outcome, treatment effects can be identified. Such an IV can frequently be thought of as exogenous encouragement or incentive to take the treatment.

Our formal discussion uses the potential outcome framework (see for instance Rubin 1974). Denote by $D(z)$ the potential treatment state that would occur if the instrument Z was exogenously set to value z. We assume a binary IV, i.e. $Z \in \{1, 0\}$, while Section 2.3 extends the framework to multivalued Z. Furthermore, denote by $Y(z, d)$ the potential outcome when setting Z and D to $z, d \in \{1, 0\}$. Importantly, $Y(z, d)$ simplifies to $Y(d)$ if Z does not directly affect Y other than through D, which is crucial in the IV context (see Assumption 1 below). In this case, the potential outcome is only a function of the treatment, as it does not depend on z after setting d. As an example, consider the experimental evaluation of a training program in which Z and D denote the randomized assignment of and the actual participation in the training, respectively. $D(1)$ and $D(0)$ denote the potential participation states when randomized into or out of the training. Similarly, $Y(1)$ and $Y(0)$ denote the potential outcomes, e.g. employment states, when participating and not participating in the training. Defining the potential outcomes as functions of training only implicitly assumes that randomized assignment does not directly affect employment other than through training participation. For each subject, only one of the two potential outcomes and treatments are observed, because the observed variables are defined as $Y = D \cdot Y(1) + (1 - D) \cdot Y(0)$ and $D = Z \cdot D(1) + (1 - Z) \cdot D(0)$.

Table 1: Definition of Types.

Types (T)	**$D(1)$**	**$D(0)$**	**Notion**
a	1	1	Always takers
c	1	0	Compliers
d	0	1	Defiers
n	0	0	Never takers

Even without any assumptions, the population can be split into four treatment compliance types (denoted by $T \in \{a, c, d, n\}$) defined by the joint potential treatment states under $z = 1$ and $z = 0$, see Angrist, Imbens, and Rubin (1996). As shown in Table 1, the compliers (c: $D(1) = 1, D(0) = 0$) react on the randomization as intended by the researcher and participate in the training when $z = 1$, while abstaining from it when $z = 0$. For the remaining three types, $D(z) \neq z$ for either $z = 1$, or $z = 0$, or both: The always takers (a: $D(1) = 1$, $D(0) = 1$) always take the training irrespectively of the IV, the never takers (n: $D(1) = 0, D(0) = 0$) are never treated, and the defiers (d: $D(0) = 1$, $D(0) = 1$) react counter-intuitively to randomization by participating in the treatment when randomized out, but not participating when randomized in. As either $D(1)$ or $D(0)$ remains unknown in the data, so does any subject's type, which is a function of both potential treatment states. This implies that any subject with a particular observed combination of the treatment and the instrument may belong to one of two types, see Table 2.

Table 2: Observed Subgroups and Types.

Observed values of Z and D	**Potential types T**
$\{Z = 1, D = 1\}$	belongs either to a or to c
$\{Z = 1, D = 0\}$	belongs either to d or to n
$\{Z = 0, D = 1\}$	belongs either to a or to d
$\{Z = 0, D = 0\}$	belongs either to c or to n

Therefore, the observable mean differences $E(Y|D = 1) - E(Y|D = 0)$ or $E(Y|D = 1, Z = z) - E(Y|D = 0, Z = z)$ (for $z \in \{1, 0\}$) do generally not correspond to any causal effect because the mixtures of types differ across D or (D, Z). The reason is that types generally have different distributions of unobservables which may confound the treatment and outcome. To convey the intuition, we consider the following nonparametric IV model:

$$Y = \phi(D, U), \quad D = \eta(Z, V), \tag{1}$$

where $\phi(\cdot)$ and $\eta(\cdot)$ denote general functions, while U and V are the unobserved terms (possibly scalars or vectors). The unobservables U and V may be arbitrarily associated with each other, thus causing the treatment to be endogenous. Potential outcomes and treatment states are obtained by exogenously setting the treatment and the IV to particular values d and z, where "exogenously setting" means manipulating D and Z without changing the values of U and V:

$$Y(1) = \phi(1, U), \quad Y(0) = \phi(0, U), \quad D(1) = \eta(1, V), \quad D(0) = \eta(0, V).$$

As $D(1) = \eta(1, V)$ and $D(0) = \eta(0, V)$ differ across types, i.e. they have different potential treatments for same values of z, the distribution of V must necessarily differ across types (as D is a function of Z and V only). Therefore, U also differs across types if it is associated with V. This can be illustrated by the following parametric model, a special case of (1):

$$Y = \beta D + U, \quad D = I\{\delta Z \geq V\}, \tag{2}$$

where β and δ are slope coefficients that are assumed to be constant across individuals. $I\{\cdot\}$ denotes the indicator function which is equal to one if its argument is satisfied and zero otherwise. Furthermore, U and V are assumed to be scalars for simplicity. For the compliers, $D(1) = I\{\delta \geq V\} = 1, D(0) = I\{0 \geq V\} = 0$, so that the distribution of V satisfies $\delta \geq V > 0$. Among always takers, however, $D(1) = I\{\delta \geq V\} = 1, D(0) = I\{0 \geq V\} = 1$, so that $V \leq 0$. Consequently, unless U and V are independent, the treatment and the outcome are confounded. Treatment effects can therefore only be identified under additional assumptions on Z, as outlined below.

By the parametric model in (2), treatment effects are homogeneous due to a constant slope coefficient in the outcome equation and additive separability of D and U. $\beta = Y(1) - Y(0)$ is constant across individuals even if $U = Y(0)$ is not. In empirical applications, however, the impact of D on Y is likely heterogenous through dependence on unobserved factors. Imbens and Angrist (1994) therefore postulate the identifying assumptions for nonparametric IV models like (1) that permit effect heterogeneity, with the caveat that effects can generally only be obtained for the subpopulation of compliers. The assumptions impose IV validity (IV exogeneity and an exclusion restriction), weak monotonicity of the treatment in the instrument, and IV relevance (for the treatment). Formally, the first assumption can be stated as follows:

Assumption 1

(IV validity). *(i)* $Z\perp(D(1),D(0),Y(1,1),Y(1,0),Y(0,1),Y(0,0))$ *and (ii)* $Y(1,d) = Y(0,d) = Y(d)$ *for* $d \in \{1, 0\}$

The symbol "$\perp$" denotes independence. Assumption 1 consists of two conditions. First, the IV is as good as random and unrelated with factors affecting the treatment and/or outcome, implying that $(U, V) \perp Z$ holds in model (1). Therefore, not only the potential outcomes/treatment states, but also the types (defined by the joint potential treatment states) are independent of the instrument. Second, Z must not have a direct effect on Y other than through D, i.e. satisfy an exclusion restriction. This holds in (1) and (2), because Z does not enter the equation of Y as an explanatory variable. We will henceforth often implicitly impose the exclusion restriction and index potential outcomes by the treatment only.

Concerning the plausibility of Assumption 1, note that in a successfully conducted experiment, the randomness of Z holds by construction. The exclusion restriction holds if mere assignment for instance to a training does not have a direct effect on the outcome, e.g. through increased motivation or frustration due to being (not) offered the training. While Assumption 1 is plausible e.g. in a medical trial where individuals in the control group receive placebo treatments, it might be less so in so-called quasi-experimental settings. Taking the estimation of the returns to education (D) as example, Angrist and Krueger (1991) suggest using quarter of birth as IV (Z). It is related to years of education through regulations about school starting age, but is arguably not driven by factors also affecting income and does not have a direct effect on income (Y). However, Bound, Jaeger, and Baker (1995) contest Assumption 1 and present evidence that seasonal birth patterns are related to family income, health, and school attendance, all of which may affect income. Buckles and Hungerman (2013) document differences in maternal characteristics for births throughout the year. Scrutinizing IVs appearing plausible at a first glance is therefore in order, in particular, when not randomly assigned by the researcher.

When aiming for a mean effect like the LATE (see (15) below), full independence between Z and $Y(z, d)$, see part (i) of Assumption 1, can be replaced by the weaker mean independence: $E(Y(z,d)|T = t, Z = 1) = E(Y(z,d)|T = t, Z = 0) = E(Y(z,d)|T = t)$ for $d \in \{0, 1\}$ and $t \in \{a, c, d, n\}$. Likewise, the exclusion restriction of part (ii) can be relaxed to a mean version: $E(Y(1,d)|T = t) = E(Y(0,d)|T = t) = E(Y(d)|T = t)$. Assumption 1 is, however, required for distributional features like quantile treatment effects. From a practical perspective, this distinction is often less relevant, as setups in which mean independence holds, but full independence does not, might seem odd. For instance, if one assumes that an IV is mean independent of the potential hourly wage, it seems reasonable that it is also mean independent of the log of potential hourly wage. As the latter is a nonlinear transformation of the original potential outcome, this implies independence also with respect to higher moments. Therefore, strengthening mean to full independence often comes with little costs in terms of credibility.

Assumption 2

(Monotonicity). $\Pr(D(1) \geq D(0)) = 1$

Assumption 2 says that the potential treatment state of any individual does not decrease in the IV. Alternatively to imposing weak positive monotonicity, one could also impose weak negative monotonicity ($\Pr(D(1) \leq D(0)) = 1$). However, the latter case is omitted, because it is symmetric in the sense that negative can be turned into positive monotonicity by recoding the instrument as $1 - Z$. Assumption 2 rules out the existence of defiers (type $T = d$), because for the latter group, $D(1) < D(0)$. The population therefore only consists of always takers, never takes, and compliers. This condition is implicit in parametric models like (2), where δ is a constant so that the effect of Z is homogeneous and V is a scalar unobservable.

Furthermore, Assumption 2 is satisfied by construction in randomized experiments with so-called one-sided non-compliance (see Bloom 1984): if no subject randomized out of a job training can manage to "sneak into" the training, then $\Pr(D(0) = 1) = 0$ such that defiers as well always takers do not exist. Even in many field experiments where $\Pr(D(0) = 1) > 0$, the presence of defiers appears implausible as it would imply counter-intuitive behavior to the randomization protocol. In several quasi-experimental settings, however, the assumption might be disputable. Reconsidering the quarter of birth instrument, positive monotonicity appears plausible in the US context at a first glance. Arguably, among students entering school in the same year, those who are born in an earlier quarter can drop out after less years of completed education at the age of 16 when compulsory schooling ends than those born later, in particular after the end of the academic year. However, strategic postponement of school entry due to redshirting or unobserved school policies may reverse the relation of education and quarter of birth for some individuals (defiers), see Aliprantis (2012), Barua and Lang (2009), and Klein (2010). Assumption 2 therefore needs to be scrutinized with similar care as Assumption 1.

Assumption 3 imposes the existence of a first stage effect of the IV on the treatment.

Assumption 3

(First stage). $E(D|Z=1) - E(D|Z=0) \neq 0$

Under Assumption 1 and Assumption 2, $E(D|Z=1) - E(D|Z=0) \neq 0$ implies the existence of compliers (type $T = c$) in the population, namely $\Pr(D(1) > D(0)) > 0$, see equation (14) further below. In our parametric model, this is satisfied if δ is positive and sufficiently large to shift the treatment decision at least for a subpopulation when switching from $z = 0$ to $z = 1$.

In seminal work, Vytlacil (2002) shows that Assumption 1 – Assumption 3 correspond to a particular nonparametric IV model (1) with the following threshold crossing selection equation

$$D = 1(\mu(Z) \geq V), \tag{3}$$

where V is a scalar (index of) unobservable(s) and $\mu(Z)$ is a nontrivial function of Z.[1]

2.2 Identification under a Binary Treatment and Instrument

To demonstrate how Assumption 1–Assumption 3 permit identifying the LATE, LQTE, and the potential outcome distributions, we introduce further notation that heavily borrows from Kitagawa (2009). Let $f(y, D = d|Z = z)$ denote the joint density of the observed outcome and $D = d$ conditional on $Z = z$ for $d, z \in \{1, 0\}$. Denote by $f(y(z,d), T = t|Z = z)$ the unobserved joint density of the potential outcome and type t conditional on $Z = z$, where $t \in \{a, c, d, n\}$. In the absence of Assumption 1–Assumption 3, it follows from Table 2 that any observed joint density is a function of the potential outcomes of two types conditional on Z. Therefore, the subsequent relationships of observed and unobserved joint densities hold for all y in the support of Y:

$$f(y, D = 1|Z = 1) = f(y(1,1), T = c|Z = 1) + f(y(1,1), T = a|Z = 1), \tag{4}$$

$$f(y, D = 1|Z = 0) = f(y(0,1), T = d|Z = 0) + f(y(0,1), T = a|Z = 0), \tag{5}$$

$$f(y, D = 0|Z = 1) = f(y(1,0), T = d|Z = 1) + f(y(1,0), T = n|Z = 1), \tag{6}$$

$$f(y, D = 0|Z = 0) = f(y(0,0), T = c|Z = 0) + f(y(0,0), T = n|Z = 0). \tag{7}$$

Under Assumption 1, $f(y(z,d), T = t|Z = z)$ simplifies to $f(y(d), T = t)$ for any type and $d, z \in \{1, 0\}$, because (i) IV independence of the potential treatments/outcomes allows dropping conditioning on Z and (ii) the exclusion restriction implies $Y(z, d) = Y(d)$. Under Assumption 2, $f(y(1), T = d)$ and $f(y(0), T = d)$ are zero. Therefore, equations (4) to (7) become

$$f(y, D = 1|Z = 1) = f(y(1), T = c) + f(y(1), T = a), \tag{8}$$

$$f(y, D = 1|Z = 0) = f(y(1), T = a), \tag{9}$$

$$f(y, D=0|Z=1) = f(y(0), T=n), \tag{10}$$

$$f(y, D=0|Z=0) = f(y(0), T=c) + f(y(0), T=n). \tag{11}$$

By Assumption 3, $f(y(0), T=c)$ and $f(y(1), T=c)$ are nonzero for at least some values $(y(0), y(1))$ in the support of $(Y(0), Y(1))$. Subtracting (9) from (8) and (10) from (11) yields the joint densities of the compliers under treatment and non-treatment:

$$f(y, D=1|Z=1) - f(y, D=1|Z=0) = f(y(1), T=c), \tag{12}$$

$$f(y, D=0|Z=0) - f(y, D=0|Z=1) = f(y(0), T=c). \tag{13}$$

To obtain the LATE, note that $\int f(y(d), T=c)dy = \pi_c$, where $\pi_c = \Pr(T=c)$ denotes the share of compliers in the population. More generally, $\pi_t = \Pr(T=t)$ will henceforth denote the share of type t. Therefore, π_c is identified by

$$\begin{aligned} \pi_c &= \int [f(y, D=1|Z=1) - f(y, D=1|Z=0)]dy \\ &= \Pr(D=1|Z=1) - \Pr(D=1|Z=0) = E(D|Z=1) - E(D|Z=0). \end{aligned} \tag{14}$$

Furthermore, $\int yf(y(d), T=c)dy = \int yf(y(d)|T=c)\pi_c dy = E[Y(d)|T=c] \cdot \pi_c$ implies that

$$\begin{aligned} &E[Y(1) - Y(0)|T=c] \cdot \pi_c \\ &\quad = \int y\{[f(y, D=1|Z=1) - f(y, D=1|Z=0)] - [f(y, D=0|Z=0) - f(y, D=0|Z=1)]\}dy \\ &\quad = \int y[f(y|Z=1) - f(y|Z=0)]dy = E(Y|Z=1) - E(Y|Z=0), \end{aligned}$$

which is the intention-to-treat effect (ITT). The latter generally deviates from the ATE because it does not comprise the effects on the always and never takers. By scaling the ITT by the share of compliers we obtain the standard identification result for the LATE, denote by Δ_c:

$$\frac{E(Y|Z=1) - E(Y|Z=0)}{E(D|Z=1) - E(D|Z=0)} = E[Y(1) - Y(0)|T=c] = \Delta_c. \tag{15}$$

That is, the so-called Wald estimand, which in the binary treatment and instrument case corresponds to the probability limit of TSLS,[2] identifies the LATE. It is worth noting that under one-sided noncompliance, the LATE simplifies to $\frac{E(Y|Z=1)-E(Y|Z=0)}{E(D|Z=1)}$ and coincides with the average treatment effect on the treated (ATT), $\Delta_{D=1} = E(Y(1) - Y(0)|D=1)$:

$$\begin{aligned} \Delta_c &= E(Y(1) - Y(0)|D(1)=1, D(0)=0) = E(Y(1) - Y(0)|D(1)=1) \\ &= E(Y(1) - Y(0)|D(1)=1, Z=1) = E(Y(1) - Y(0)|D=1, Z=1) = E(Y(1) - Y(0)|D=1). \end{aligned}$$

The second equality follows from $\Pr(D(0)=1) = 0$ such that $D(1) = 1$ implies $T = c$, the third from Assumption 1, the fourth from the definition of potential treatments, and the fifth from $\Pr(D(0)=1) = 0 \Rightarrow \Pr(D=1|Z=0) = 0$ such that $D=1 \Rightarrow D=1, Z=1$.

Also the density functions of the potential outcomes among compliers are identified, see Imbens and Rubin (1997). By (12), (13), and (14),

$$\begin{aligned} f(y(1)|T=c) &= \frac{f(y, D=1|Z=1) - f(y, D=1|Z=0)}{\Pr(D=1|Z=1) - \Pr(D=1|Z=0)} \\ &= \frac{f(y|D=1, Z=1) \cdot \Pr(D=1|Z=1) - f(y|D=1, Z=0) \cdot \Pr(D=1|Z=0)}{\Pr(D=1|Z=1) - \Pr(D=1|Z=0)}. \\ f(y(0)|T=c) &= \frac{f(y, D=0|Z=0) - f(y, D=0|Z=1)}{\Pr(D=0|Z=0) - \Pr(D=0|Z=1)} \\ &= \frac{f(y|D=0, Z=0) \cdot \Pr(D=0|Z=0) - f(y|D=0, Z=1) \cdot \Pr(D=0|Z=1)}{\Pr(D=0|Z=0) - \Pr(D=0|Z=1)}. \end{aligned}$$

The mean potential outcomes among compliers correspond to the following expressions, see also Imbens and Rubin (1997) and Abadie (2002):

$$\begin{aligned} E(Y(1)|T=c) &= \frac{\int y\{f(y,D=1|Z=1)-f(y,D=1|Z=0)\}dy}{\Pr(D=1|Z=1)-\Pr(D=1|Z=0)} \\ &= \frac{E(Y\cdot D|Z=1)-E(Y\cdot D|Z=0)}{E(D|Z=1)-E(D|Z=0)}. \end{aligned} \tag{16}$$

$$\begin{aligned} E(Y(0)|T=c) &= \frac{\int y\{f(y,D=0|Z=0)-f(y,D=0|Z=1)\}dy}{\Pr(D=0|Z=0)-\Pr(D=0|Z=1)} \\ &= \frac{E(Y\cdot(1-D)|Z=1)-E(Y\cdot(1-D)|Z=0)}{E(1-D|Z=1)-E(1-D|Z=0)}. \end{aligned}$$

$E(Y(1)|T=c)$ can be consistently estimated by TSLS when using Z as instrument in a regression of $Y \cdot D$ on a constant and D, where the coefficient on the latter gives the estimate. An estimate of $E(Y(0)|T=c)$ is obtained from a TSLS regression of $Y \cdot (1-D)$ on $(1-D)$.

As shown in Lemma 2.1 of Abadie (2002), the identification results (16) not only hold with respect to Y, but also for any function of the outcome, denoted by $h(y)$, with a finite first moment. As an important case, setting $h(y) = 1(Y \le y)$, with y being some value on the real line, allows identifying cumulative distribution functions (cdf) of potential outcomes:

$$F_{Y(1)|T=c}(y) = \frac{E(1(Y\le y)\cdot D|Z=1)-E(1(Y\le y)\cdot D|Z=0)}{E(D|Z=1)-E(D|Z=0)}, \tag{17}$$

$$F_{Y(0)|T=c}(y) = \frac{E(1(Y\le y)\cdot(1-D)|Z=1)-E(1(Y\le y)\cdot(1-D)|Z=0)}{E(1-D|Z=1)-E(1-D|Z=0)}.$$

Estimation can proceed by TSLS, regressing $1(Y \le y) \cdot D$ on D and $1(Y \le y) \cdot (1-D)$ on $(1-D)$. Quantiles of the potential outcomes of compliers are obtained by inverting the cdfs:

$$Q_{Y(d)|T=c}(\tau) = \left\{\inf_y \Pr(Y(d) \le y|T=c) \ge \tau\right\} = F^{-1}_{Y(d)|T=c}(\tau), \tag{18}$$

where $\tau \in (0,1)$ is the rank in the potential outcome distribution under $D = d$. This allows for defining the local quantile treatment effect (LQTE) at the τth quantile, which corresponds to

$$\Delta_c(\tau) = Q_{Y(1)|T=c}(\tau) - Q_{Y(0)|T=c}(\tau). \tag{19}$$

Estimation can be performed by inverting the empirical potential outcome cdfs. Under standard regularity conditions estimation is consistent and asymptotically normal if the densities of the potential outcomes among compliers are strictly positive: $f(y(d)|T=c) > 0$ for $d \in \{0,1\}$.

2.3 Multivalued Instruments and Marginal Treatment Effects

This section considers extensions to nonbinary IVs while maintaining a binary treatment. If the IV is multivalued one can identify a LATE with respect to any pair of values (z'', z') satisfying Assumption 1–Assumption 3. Instead of identifying many pairwise effects, one might be interested in the effect for the largest possible complier population. By defining monotonicity with respect to the treatment propensity score $p(z) = \Pr(D=1|Z=z)$, this is achieved by finding the two instrument values that minimize and maximize $p(z)$, $(z_{\min}, z_{\max})$, and evaluating the LATE at these values. The propensity score (rather than Z itself) might be used as IV for identification:

$$\Delta_c(p(z_{\min}), p(z_{\max})) = \frac{E(Y|p(Z)=p(z_{\max}))-E(Y|p(Z)=p(z_{\min}))}{E(D|p(Z)=p(z_{\max}))-E(D|p(Z)=p(z_{\min}))}.$$

This strategy is particularly useful if Z is multidimensional, as the different elements in Z can then be straightforwardly collapsed into a single instrument by using $p(Z)$.

If the IV(s) is/are continuous, a continuum of effects is identified. See Heckman and Vytlacil (2001b, 2005), who call the resulting parameter based on an infinitesimal change in the IV the marginal treatment effect (MTE). The latter is defined as average treatment effect conditional on V, the unobserved term in the treatment model (1):

$$\Delta(v) = E(Y(1) - Y(0)|V = v). \tag{20}$$

Assume that V represents the (unobserved) cost or disutility of treatment. The MTE is the average effect among persons being indifferent between treatment or not if exogenously assigned a value of Z, say z, such that $\mu(z) = v$. This follows from $D = 1(\mu(Z) \geq V)$, see (3). Any LATE (and any other average effect) can be expressed as a (density-)weighted average of MTEs. For any (z'', z') such that $p(z'') > p(z')$, a complier is someone satisfying $D(z'') = 1(\mu(z'') \geq V) = 1$ and $D(z') = 1(\mu(z') \geq V) = 0$. Put differently, compliers $c(z'', z')$ are characterized by $v' < V \leq v''$ so that $D(z'') = 1$ and $D(z') = 0$ holds. Therefore, the LATE for $T = c(z'', z')$ is

$$\begin{aligned} &E(Y(1) - Y(0)|T = c(z'',z')) = E(Y(1) - Y(0)|D(z'') = 1, D(z') = 0) \\ &\quad = E(Y(1) - Y(0)|v' < V \leq v'') = \Delta_c(v'',v') = \frac{1}{F_V(v'') - F_V(v')} \int_{v'}^{v''} \Delta(v) dF_V(v). \end{aligned}$$

V can be normalized so that the normalization (denoted by $\bar{V}$) satisfies $\bar{V} \sim \text{Uniform}[0,1]$ and corresponds to the cdf: $\bar{V} = F_V$. Normalizing is innocuous, because if $D = 1(\mu(Z) \geq V)$, then by applying a probability transformation, the model can be reparametrized so that $D = 1(\eta(Z) \geq \bar{V})$, with $\eta(Z) = F_V(\mu(Z))$. Thus, $\Delta_c(\bar{v}'',\bar{v}') = \frac{1}{\bar{v}''-\bar{v}'} \int_{\bar{v}'}^{\bar{v}''} \Delta(\bar{v}) d\bar{v}$.

The MTE can be identified by the fact that $\bar{v}'' = F_V(v'') = \Pr(D(z'') = 1) = \Pr(D = 1|Z = z'') = p(z'')$ and equivalently, $\bar{v}' = p(z')$. Therefore, the MTE is recovered pointwise by the derivative of the conditional expectation of Y with respect to $p(Z)$:

$$\Delta(\bar{V} = p(z)) = \frac{\partial E(Y|p(Z) = p(z))}{\partial p(z)}. \tag{21}$$

Heckman and Vytlacil (1999) coined the term local IV (LIV) for $\Delta(\bar{V} = p)$, a parameter even "more local" than the LATE $\Delta_c(\bar{v}'',\bar{v}')$ based on a quantifiable difference between $\bar{v}''$ and $\bar{v}'$. Note, however, that the LATE is equivalent to the LIV for $\bar{v}'' - \bar{v}'$ infinitesimally small. Using similar arguments, Carneiro and Lee (2009) extend these ideas to the identification of the QTE analogs of the MTE, the marginal quantile treatment effects (MQTE):

$$\Delta(\tau|\bar{V} = p(z)) = Q_{Y_1}(\tau|\bar{V} = p(z)) - Q_{Y_0}(\tau|\bar{V} = p(z)). \tag{22}$$

$Q_{Y_1}(\tau|\bar{V} = p(z))$ and $Q_{Y_0}(\tau|\bar{V} = p(z))$ are identified as the inverses of

$$\begin{aligned} F_{Y(1)}(y|\bar{V} = p(z)) &= F_Y(y|P(Z) = p(z), D = 1) + p(z)\frac{\partial F_Y(y|P(Z) = p(z), D = 1)}{\partial p}, \\ F_{Y(0)}(y|\bar{V} = p(z)) &= F_Y(y|P(Z) = p(z), D = 0) - (1 - p(z))\frac{\partial F_Y(y|P(Z) = p(z), D = 0)}{\partial p}. \end{aligned}$$

2.4 Multivalued Treatments

In contrast to extending binary to multivalued IVs (see Section 2.3), generalizing binary to nonbinary treatments is not straightforward. For illustration, consider a setup with a single binary instrument $Z \in \{0, 1\}$ and an ordered discrete treatment $D \in \{0, 1, \ldots, J\}$, where $J + 1$ is the number of possible treatment doses. We cannot identify causal effects for single compliance types at specific treatment values, e.g. for those increasing the treatment from 1 to 2 when the instrument switches from 0 to 1. Is is, however, possible to identify a weighted average of causal effects of unit-level increases in the treatment. Angrist and Imbens (1995) show that if $\Pr(D(1) \geq j > D(0)) > 0$ for some value j such that compliers exist at some treatment margin,

$$\frac{E(Y|Z = 1) - E(Y|Z = 0)}{E(D|Z = 1) - E(D|Z = 0)} = \sum_{j=1}^{J} w_j \cdot E(Y(j) - Y(j-1)|D(1) \geq j > D(0)), \tag{23}$$

with $w_j = \frac{\Pr(D(1) \geq j > D(0))}{\sum_{j=1}^{J} \Pr(D(1) \geq j > D(0))}$, implying that $0 \leq w_j \leq 1$ and $\sum_{j=1}^{J} w_j = 1$. Therefore, the Wald estimand equals a weighted average of effects of unit changes in the treatment on heterogeneous complier groups defined by different margins of the potential treatments. However, the various treatment effects based on unit changes, $E(Y(j) - Y(j-1)|D(1) \geq j > D(0))$, remain unidentified. Furthermore, complier groups might be overlapping. Some individuals could satisfy both $(D(1) \geq j > D(0))$ and $(D(1) \geq j + 1 > D(0))$ for some j and therefore be accounted multiple times. This arguably compromises the interpretability of the effect.[3] Angrist and Imbens (1995) show that similar results hold in setups with multiple instruments and covariates. While this strategy yields weighted LATEs, it cannot be used for LQTEs, which rely on separately identifying and inverting marginal distributions of potential outcomes.

In contrast to an ordered treatment, Behaghel, Crépon, and Gurgand (2013) consider multiple unordered treatments that are mutually exclusive, which is equivalent to the case of a single treatment with multiple, albeit unordered values. Under Assumption 1 and a particular monotonicity assumption tailored to a three-valued treatment and instrument $(D, Z \in \{0, 1, 2\})$, LATEs among the two complier populations $c_1 : D(1) = 1, D(0) = 0$ and $c_2 : D(2) = 2, D(0) = 0$ are identified.[4] Kirkeboen, Leuven, and Mogstad (2016) obtain identification under a somewhat weaker monotonicity assumption by exploiting information about individuals' preferred treatments and next-best treatment alternatives. Heckman and Pinto (2018) consider an unordered monotonicity assumption. It requires for any specific value of the unordered treatment that if some subjects move into (out of) the respective value when the instrument is switched, then no subjects can at the same time move out of (into) that value. Imposing conditional IV validity in the spirit of Assumption 4, Hull (2015) shows under a modified monotonicity assumption for a three-valued treatment that LATEs can be obtained from a binary instrument if (i) compliance is heterogeneous and (ii) LATEs are homogeneous in observables X. Lee and Salanie (2015) assume that any treatment value is a measurable function of some threshold-crossing models and sufficiently many continuous instruments are available, but require no classical monotonicity assumption.

3 Treatment Evaluation with Covariates

Section 3.1 presents the identifying assumptions in the presence of covariates. Sections 3.2, 3.3, and 3.4 consider local, quantile, and marginal treatment effects. Section 3.5 discusses the identification of general functionals among compliers.

3.1 Identifying Assumptions

It may not appear credible that an instrument satisfies Assumption 1–Assumption 3 unconditionally, i.e. without controlling for further covariates. This is commonly the case in observational data in which the instrument is typically not explicitly randomized like in an experiment. As an example, consider the study of Card (1995), who evaluates the returns to education using the US National Longitudinal Survey of Young Men. Geographic proximity to college serves as instrument for the potentially endogenous education decision. Proximity should induce some individuals to strive for a college degree who would otherwise not, for instance due to costs associated with not living at home. However, the instrument might be correlated with factors like local labor market conditions or family background. These factors might be related to the earnings outcome, implying a violation of Assumption 1. For this reason, Card (1995) includes a range of control variables, e.g. parents' education, ethnicity, urbanity, and geographic region.

We subsequently reconsider the binary instrument and treatment case of Section 2.1. However, we impose conditional IV assumptions (see for instance Abadie 2003), implying that the IV assumptions only hold when controlling for a vector of observed covariates denoted by X.

Assumption 4

(Conditional IV validity). *(i)* $Z \perp (D(1), D(0), Y(1,1), Y(1,0), Y(0,1), Y(0,0))|X$ *and (ii)* $\Pr(Y(1,d) = Y(0,d) = Y(d)|X) = 1$ *for* $d \in \{1, 0\}$

Assumption 5

(Conditional monotonicity). $\Pr(D(1) \geq D(0)|X) = 1$

Assumption 6

(Conditional first stage). $E(D|Z = 1, X) - E(D|Z = 0, X) \neq 0$

Assumption 4 is weaker than Assumption 1, because IV validity now is only required to hold among units with the same values of X rather than unconditionally. Assumption 5 requires that defiers do not exist for every value of X. Theoretically, one could construct cases where defiers exist unconditionally (such that $\Pr(D(1) \geq D(0)) = 1$ as stated in Assumption 2 does not hold), but not after conditioning on X, for instance if Z affected X positively and X affected D (sufficiently strongly) negatively. Assumption 6 states that the first stage is non-zero for every value of X in its support, which is stronger than Assumption 3. Under Assumption 4 and Assumption 5, this implies $\Pr(D(1) > D(0)|X) > 0$, which allows identifying the conditional LATE or LQTE almost everywhere, see Sections 3.2 and 3.3. In contrast, $\Pr(D(1) > D(0)) > 0$ would suffice if one was only interested in the (unconditional) LATE and LQTE.

Assumption 7
(Common support). $0 < \Pr(Z = 1|X) < 1$

Assumption 7 is a common support restriction requiring that no value of X perfectly predicts instrument assignment. Otherwise, no comparable units (in terms of X) across IV states $Z = 1$ and $Z = 0$ exist at some values of X so that identification breaks down.

Similar to (1), we briefly consider a general IV model that now includes X:

$$Y = \phi(D, X, U), \quad Y(d) = \phi(d, X, U), \quad D = \delta(Z, X, V), \quad D(z) = \delta(z, X, V). \tag{24}$$

Assumption 4 implies that $(U, V) \perp Z|X$. Furthermore, under Assumption 5–Assumption 6, D can also be represented as $D = 1(\mu(Z, X) \geq V)$, see Vytlacil (2002).

3.2 LATE

Under Assumption 4–Assumption 7, the conditional LATE given $X = x$ is identified by

$$\Delta_c(x) = E(Y(1) - Y(0)|T = c, X = x) = \frac{E(Y|Z = 1, X = x) - E(Y|Z = 0, X = x)}{E(D|Z = 1, X = x) - E(D|Z = 0, X = x)}, \tag{25}$$

see for instance Heckman (1997). Nonparametric estimation of $\Delta_c(x)$ suffers from the curse of dimensionality when X is high dimensional. To overcome this problem, one may either impose parametric restrictions on the conditional means $E(Y|Z = z, X = x)$ and $E(D|Z = z, X = x)$ for $z \in \{0, 1\}$ as in Tan (2006), or employ the weighting result by Abadie (2003) to construct semiparametric weighted regression estimates, see Section 3.5.

While identification of the conditional LATE permits investigating effect heterogeneity with respect to observable covariates, the (unconditional) LATE is frequently the main parameter of interest, also under conditional IV assumptions. It is obtained as a weighted average of conditional LATEs among compliers, i.e. by integrating over the distribution of X given $T = c$: $\Delta_c = \int \Delta_c(x) dF_{X|T=c}(x)$. Frölich (2007) shows that the LATE can also be represented by:

$$\Delta_c = \frac{\int \{E(Y|Z = 1, X = x) - E(Y|Z = 0, X = x)\} dF_X(x)}{\int \{E(D|Z = 1, X = x) - E(D|Z = 0, X = x)\} dF_X(x)}. \tag{26}$$

By noting that $\int E(Y|Z = 1, X = x) dF_X(x) = \int (1/\pi(x)) E(Y \cdot Z|X = x) dF_X(x) = E(Y \cdot Z/\pi(X))$, where $\pi(x) = \Pr(Z = 1|X = x)$ is the instrument propensity score, one can also obtain a weighting-based expression, see Tan (2006) and Frölich (2007):

$$\Delta_c = \frac{E[Y \cdot Z/\pi(X) - Y \cdot (1 - Z)/(1 - \pi(X))]}{E[D \cdot Z/\pi(X) - D \cdot (1 - Z)/(1 - \pi(X))]}. \tag{27}$$

Using a result of Rosenbaum and Rubin (1983) which shows that for identification, controlling for the propensity score is as good as controlling for X, a third representation of the LATE is

$$\Delta_c = \frac{\int \{E(Y|Z = 1, \pi(X) = p) - E(Y|Z = 0, \pi(X) = p)\} dF_\pi(p)}{\int \{E(D|Z = 1, \pi(X) = p) - E(D|Z = 0, \pi(X) = p)\} dF_\pi(p)}. \tag{28}$$

This has the advantage that $\pi(X)$ is one-dimensional, no matter of which dimension X is. The LATE can thus be estimated as the ratio of two propensity score matching estimators with Z being the "treatment" and either Y (numerator) or D (denominator) being the "outcome."

Several analog estimators have been proposed based on (26) and (27). Frölich (2007) analyzes nonparametric matching- and (local polynomial and series) regression-based estimation of (26). Donald, Hsu, and Lieli (2014a, 2014b) propose nonparametric inverse probability weighted estimators of (27), using series logit and local polynomial regression-based estimation of the instrument propensity score. All estimators are $\sqrt{n}$-consistent and asymptotically normal under specific regularity conditions, as fully nonparametric estimation of the unconditional LATE involves averaging over conditional LATEs and does therefore not give rise to the curse of dimensionality. Hong and Nekipelov (2010) derive semiparametric efficiency bounds for the estimation of nonlinear LATE models and propose efficient estimators.

Parametric estimation strategies for the unconditional LATE are outlined in Tan (2006) and Uysal (2011), who both propose estimators that rely on parametric models for the propensity scores and conditional expectations of the outcomes. To guard against misspecification, they consider so-called doubly-robust (DR) estimators. DR estimators are consistent if either the propensity score, the conditional expectations, or both are correctly specified.

Finally, Belloni et al. (2017) and Chernozhukov et al. (2017) consider estimation of the unconditional LATE in data-rich environments where the number of potential control variables may be much larger than the number of observations. Belloni et al. (2017) propose estimators based on LASSO and post-LASSO. A crucial assumption for valid inference is "approximate sparsity," implying that a relatively small number of covariates suffices for capturing the relationship of X with the outcome, treatment, and instrument up to a small approximation error. Given that sparsity holds, the advantage of the method of Belloni et al. (2017) is that it selects covariates (and possibly interaction terms and higher order terms thereof similar to series estimation) in a data-driven way rather than ad hoc. Chernozhukov et al. (2017) offer a general approach to machine learning-based LATE estimation in the presence of high dimensional control variables based on orthogonalized moment conditions and sample splitting (or cross-fitting). In addition to LASSO, further machine learning methods like random forests, neural nets, boosting, and hybrids thereof might be used to control for X.

When the IV assumptions hold conditionally on X, the LATE among all compliers is different from the local average treatment effect among treated compliers (LATT), which is considered for example by Hong and Nekipelov (2010). The reason is that the distribution of X generally differs across treatment states. By appropriate reweighting of the previous identification results, also the LATT is identified. One approach is weighting observations in the denominator and numerator of expression (27) by $\pi(X)/\Pr(Z=1)$, see Donald, Hsu, and Lieli (2014b), yielding:

$$\Delta_{c,D=1} = \frac{E\left(Y\cdot Z - Y\cdot(1-Z)\cdot\pi(X)/(1-\pi(X))\right)}{E\left(D\cdot Z - D\cdot(1-Z)\cdot\pi(X)/(1-\pi(X))\right)}. \tag{29}$$

Note that in the case of one-sided non-compliance given X, $\Pr(D(0)=1|X)=0$, the LATE does not correspond to the ATT under Assumption 4–Assumption 7 (in contrast to Assumption 1–Assumption 3). Frölich and Melly (2013a) show that in this case, the ATT is identified by

$$\Delta_{D=1} = \frac{E(Y) - \int E(Y|Z=0,X=x)dF_X(x)}{\Pr(D=1)} = \frac{1}{\Pr(D=1)}E\left[Y\cdot\left(D-(1-D)\cdot\frac{\pi(X)-Z}{1-\pi(X)}\right)\right].$$

3.3 LQTE

As for the LATE, one may define either conditional (given X) or unconditional LQTEs in the presence of covariates. This distinction is important because of the definition of quantiles. Suppose that we are interested in the relationship between education and wages. The unconditional 0.9 quantile of the wage distribution refers to high wage workers who typically have many years of schooling, whereas the 0.9 quantile of the wage distribution conditional on schooling refers to the high wage earners within an education class who will not necessarily be high overall earners; See Frölich and Melly (2013b), who provide a more detailed discussion about the difference between conditional and unconditional LQTEs. Abadie, Angrist, and Imbens (2002) consider estimation of the conditional LQTE. Assuming that the conditional quantile function for the compliers satisfies

$$Q_{Y|D,X,T=c}(\tau) = \alpha_c(\tau)D + X'\beta_c(\tau), \tag{30}$$

they show that conditional LQTE, $\Delta_c(\tau|x)$, is identified by $\alpha_c(\tau)$, the coefficient on D in a weighted quantile regression objective function:

$$(\alpha_c(\tau), \beta_c(\tau)) = \arg\min_{a,b} E\left(\kappa \cdot \rho_\tau(Y - aD - X'b)\right). \tag{31}$$

κ, which is defined in expression (34) in Section 3.5 below, is a weighting function that allows identifying functionals for compliers. Note that among the population of compliers, outcome comparisons by D conditional on X as in (30) have a causal interpretation. This follows from Assumption 4 and the fact that compliers satisfy $D = Z$:

$$Z \perp (D(1), D(0), Y(1), Y(0))|X \Rightarrow Z \perp (Y(1), Y(0))|X, T = c \Rightarrow D \perp Y(1), Y(0)|X, T = c.$$

Although the population objective function (31) is globally convex, its sample counterpart is typically not as κ is negative when $D \neq Z$. Abadie, Angrist, and Imbens (2002) therefore suggest replacing the κ-weights by their projections on (Y, D, X), which are guaranteed to be positive. Estimation consists of two steps: (i) nonparametric power series estimation of the weights and (ii) a weighted quantile regression using the estimated weights from the first step. Under appropriate regularity conditions, the resulting estimators $\hat{\alpha}_c(\tau)$ and $\hat{\beta}_c(\tau)$ are $\sqrt{n}$-consistent and asymptotically normal, because the outcome equation is parametric.

Consider next the problem of identifying and estimating the unconditional LQTE when controlling for covariates. First, note that the unconditional complier cdf – in analogy to (17) combined with (26) – is identified as follows (Frölich and Melly 2013b):

$$F_{Y(1)|T=c}(y) = \frac{E(\kappa_{FM} \cdot 1(Y \leq y) \cdot D)}{E(\kappa_{FM} \cdot D)}, \text{ where } \kappa_{\mathrm{FM}} = \frac{Z - \pi(X)}{\pi(X) \cdot (1 - \pi(X))} \cdot (2D - 1). \tag{32}$$

An analogous result holds for $F_{Y(0)|T=c}(y)$ by replacing D with $1 - D$, such that the unconditional LQTE is $\Delta_c(\tau) = F^{-1}_{Y(1)|T=c}(\tau) - F^{-1}_{Y(0)|T=c}(\tau)$. Alternatively, the unconditional QTE can be obtained by weighted quantile regression:

$$(\alpha_c(\tau), \beta_c(\tau)) = \arg\min_{a,b} E\left(\kappa_{\mathrm{FM}} \cdot \rho_\tau(Y - aD - b)\right). \tag{33}$$

Finally, Frölich and Melly (2013a) show that under one-sided noncompliance, the quantile treatment effect on the treated is given by $\Delta_{D=1}(\tau) = Q_{Y|D=1}(\tau) - F^{-1}_{Y(0)|D=1}(\tau)$, where

$$F_{Y(0)|D=1}(\tau) = \frac{1}{\Pr(D = 1)} E\left(1(Y \leq q) \cdot (1 - D) \cdot \frac{\pi(X) - Z}{1 - \pi(X)}\right).$$

Hsu, Lai, and Lieli (2015) derive uniformly consistent and asymptotically Gaussian estimators based on (32) using series logit regression for propensity score estimation. Frölich and Melly (2013b) estimate (33) using local polynomial regression for propensity score estimation. Finally, Belloni et al. (2017) consider estimation of the unconditional LQTE in data-rich environments, in which the number of potential control variables may be larger than the number of observations.

3.4 Marginal Treatment Effects

In the presence of covariates, the marginal treatment effect given X, $\Delta(x, v) = E(Y(1) - Y(0)|X = x, V = v)$, is identified by LIV: $\Delta(X = x, \bar{V} = p(z,x)) = \frac{\partial E(Y|X=x, p(Z,X)=p(z,x))}{\partial p(z,x)}$, with $p(z,x) = \Pr(D = 1|Z = z, X = x)$. Assumption 4–Assumption 6 must hold for all values of $p(Z, X)$ of interest. Assumption 7 adapted to the continuous instrument $p(Z, X)$ implies that the MTE is only identified over the common support of $p(Z, X)$ across all values of X. This limits the practical feasibility of nonparametric MTE evaluation, in particular if X is high dimensional and Z is not strong or rich in support. We refer to Cornelissen et al. (2016) for an introduction and overview of different methods for estimating MTE with covariates.

Carneiro, Heckman, and Vytlacil (2011) increase identifying power by replacing Assumption 4 with:

Assumption 8
$(Z, X) \perp (D(z), D(z'), Y(1), Y(0))$ *for* z, z' *in the support of* Z

Note that $z = 1$, $z' = 0$ in the binary IV case. This restriction imposes independence between X and unobservables affecting D or Y, which is substantially stronger than Assumption 4. While observables X (e.g. education) may confound Z and (D, Y), they must not be associated with unobservables (e.g. ability) that affect (D, Y), i.e. with (V, U) in model (24). If Assumption 8 is nevertheless imposed, the MTE is identified over the unconditional support of $p(Z, X)$ such that common support across values of X is not required. Brinch, Mogstad, and Wiswall (2017) show that this property is even maintained when relaxing Assumption 8 by permitting associations between V and X, as long as X is conditionally independent of U given V.

Identification of MQTE in the presence of covariates follows from the arguments of Section 2.3 conditional on X. Carneiro and Lee (2009) propose a semiparametric estimation approach which relies on additive separability of the structural functions determining potential outcomes. In contrast, Yu (2014) proposes a strategy that not relying on separability.

3.5 General Functionals

Abadie (2003) shows that under Assumption 4–Assumption 7, it is possible to identify a broad class of functionals for the compliers, rather than merely treatment effects. For any real function $g(Y, D, X)$ with a finite first moment and weighting functions

$$\begin{aligned} \kappa_{(0)} &= (1-D)\cdot\frac{(1-Z)-(1-\pi(X))}{(1-\pi(X))\cdot\pi(X)}, \quad \kappa_{(1)} = D\cdot\frac{Z-\pi(X))}{(1-\pi(X))\cdot\pi(X)}, \\ \kappa &= \kappa_{(0)}\cdot(1-\pi(X)) + \kappa_{(1)}\cdot\pi(X) = 1-\frac{D\cdot(1-Z)}{1-\pi(X)}-\frac{(1-D)\cdot Z}{\pi(X)}, \end{aligned} \tag{34}$$

it holds that $E(g(Y,D,X)|T=c) = \frac{E(\kappa\cdot g(Y,D,X))}{E(\kappa)}$, $E(g(Y(0),X)|T=c) = \frac{E(\kappa_{(0)}\cdot g(Y,X))}{E(\kappa)}$, $E(g(Y(1),X)|T=c) = \frac{E(\kappa_{(1)}\cdot g(Y,X))}{E(\kappa)}$: the weighting functions κ, $\kappa_{(1)}$, and $\kappa_{(0)}$ allow identifying functions (e.g. regression functions) for compliers and for compliers under (non-)treatment.

Section 3.3 has presented an application of this general weighting result to evaluate the conditional LQTE. As a further application, consider the linear outcome model $Y = X'\alpha + \beta D + U, \quad E(U|X,D) = 0$, where α and β are homogeneous coefficients on the covariates and the treatment, respectively. The optimization problem is $(\alpha_c, \beta_c) = \arg\min_{a,b} E((Y - X'a - bD)^2|T = c) = \arg\min_{a,b} E(\kappa\cdot(Y - X'a - bD)^2)$. Note that division by $E(\kappa)$ is not required as it does not affect the minimization problem. β_c gives the conditional LATE, $\Delta_c(x)$, which in our linear model also corresponds to the (unconditional) LATE, Δ_c, as well as the treatment effect in the entire population. In contrast to TSLS, this approach does not require specifying a model for D, but instead relies on a model for the instrument propensity score $\pi(X)$.

4 Some Extensions

Section 4.1 presents approaches to LATE evaluation under outcome attrition, outcome non-response, or sample selection. Section 4.2 discusses methods for dealing with measurement errors in the treatment or the instrument. Section 4.3 considers identifying the effects of dynamic, i.e. sequentially assigned, or multiple treatments. Section 4.4 is concerned with disentangling the (total) LATE into various causal mechanisms or direct and indirect effects.

4.1 LATE Evaluation under Outcome Attrition and Sample Selection

In addition to treatment endogeneity, evaluation is frequently complicated by selective attrition in the outcome. Examples are drop-out bias in follow-up surveys measuring the outcome and sample selection, for example, when the wage outcome is only observed for the working. Outcome non-response is frequently modeled by a so-called missing-at-random (MAR) restriction, which assumes conditional independence of attrition and outcomes given observed variables (e.g. Z, D, X), see for instance Rubin (1976) and Little and Rubin (1987). An alternative to MAR which is particularly tailored to the LATE framework is the so-called latent ignorability (LI) assumption of Frangakis and Rubin (1999). LI requires outcome non-response to be independent of the potential outcomes conditional on the compliance type. Furthermore, MAR and LI might be combined such that independence is assumed conditional on both observed characteristics and compliance types, see for instance

Mealli et al. (2004): $Y \perp R | Z, T, X$, where R is a binary indicator for outcome Y. This condition is equivalent to $Y \perp R | Z, D, T, X$ as Z and T perfectly determine D. Frölich and Huber (2014) extend LATE identification under MAR and both MAR and LI to dynamic non-response models with multiple outcome periods.

A shortcoming of LI (and MAR) is that outcome non-response must be related in a rather restrictive way to unobservables affecting the outcome: The compliance type is assumed to be a sufficient statistic for the association between response and unobservables, at least conditional on observed variables. So-called non-ignorable non-response models do not impose such restrictions on the relation of R and, for instance, U in (24). However, without a second instrument for R, the LATE is only identified under tight structural assumptions, see for instance Zhang, Rubin, and Mealli (2009) and Frumento et al. (2012). In contrast, Fricke et al. (2015) discuss nonparametric LATE identification when a continuous instrument for R is available in addition to the binary instrument for D and present an application in which the instruments are independently randomized. Chen and Flores (2015) do not consider instruments, LI, or MAR with respect to response, but partially identify the LATE based on imposing monotonicity of R in D among compliers. Furthermore, they assume a particular order of mean potential outcomes under specific treatments across various subpopulations defined in terms of compliance and response.

4.2 Measurement Error in the Treatment or Instrument

Ura (2016) discusses LATE evaluation when the treatment is measured with error, i.e. misclassified. While point identification is generally lost, the study provides upper and lower bounds when Assumption 1–Assumption 3 are satisfied with respect to the true treatment. Ura (2016) clarifies that the Wald estimand may generally lie outside the identified set and is only included in the latter if conditions (39) in Section 6.1 are satisfied with respect to the mismeasured treatment.

Several contributions present methods that attain point identification of the LATE in spite of treatment misclassification by imposing additional assumptions. Battistin, De Nadai, and Sianesi (2014) exploit multiple measurements of the treatment in different data sets under the condition that misclassifications in the various measurements are independent of each other conditional on the true treatment and Z. DiTraglia and Garcia-Jimeno (2016) assume additive separability of the unobservable in the treatment effect model (implying effect homogeneity conditional on covariates) and independence of the misclassification probability and Z. They also invoke mean independence of the (demeaned) unobservable, its square, and its cube of Z, which holds under Assumption 1. Yanagi (2017) assumes an exogenous variable that shifts the true treatment (in addition to Z), influences the outcome homogenously across true treatment states given Z, and does not affect misclassification given the true treatment and Z.

Chalak (2016) considers measurement error in the instrument rather than the treatment. Denoting by W and Z the mismeasured and true instrument, respectively, W is assumed to be mean independent of Y and D given Z and to satisfy an exclusion restriction, while monotonicity is not imposed. In the binary instrument case and under the satisfaction of Assumption 1–Assumption 3, W identifies the same LATE that would have been recovered under Z. For more general settings with multiple treatment and/or instrument values, Chalak (2016) shows that the Wald and LIV estimands using W identify weighted averages of LATEs or MTEs. He also discusses necessary and sufficient conditions for the weights being nonnegative.

4.3 Dynamic and Multiple Treatments

Rather than evaluating the effects of single treatments, one might be interested in the impact of several sequentially assigned (i.e. dynamic) treatments that take place at various points in time. Consider for instance the effectiveness of sequences of active labor market policies like a job application training, which is followed by an IT course and a subsidized employment program. This sequence could be compared to non-participation in any program or a different sequence of interventions. Such a dynamic treatment framework generally requires multiple instruments for each of the treatments and specific multi-period monotonicity conditions. More formally, consider a set up with two treatment periods and let D_1 and D_2 denote the first and second binary treatment. Potential outcomes $Y(d_1, d_2)$ are now defined in terms of two treatment interventions (with d_1, d_2 $\in \{1, 0\}$). Furthermore, let T_1 and T_2 denote the compliance types defined in terms of the reaction of D_1 to the first instrument Z_1 and of D_2 to the second instrument Z_2. Miquel (2002) discusses various conditions under which dynamic LATEs for specific types defined in terms of first- and second-period compliance are identified in panel data. She for instance considers the identification among compliers w.r.t. either instrument: $E\left(Y(d_1, d_2') - Y(d_1'', d_2''')|T_1 = c, T_2 = c\right)$ for $d_1, d_2', d_1'', d_2''' \in \{1, 0\}$. Miquel (2002) also shows that if only one IV is available for both treatment periods, only the effects of particular sequences can (under specific assumptions) be identified for individuals that are always or never takers in the first treatment and compliers in the second one or vice versa.

In a multiple treatment framework, various treatments are not assigned sequentially, but rather at the same point in time. At a first glance, the simultaneous availability of several binary treatments (e.g. alternative active labor market policies) constitutes a similar evaluation problem like a treatment with multiple unordered values as discussed in Section 2.4. However, one important distinction is that multiple treatments need not be mutually exclusive and in fact, one might be interested in the effect of assigning several treatments at the same time. This generally requires distinct instruments for each treatment. Blackwell (2015) considers LATE identification of separate and joint effects of two treatments in various subpopulations defined upon compliance with either of the binary instruments, namely: $E(Y(1,1) - Y(0,0)|T_1 = c, T_2 = c)$, $E(Y(1,0) - Y(0,0)|T_1 = c, T_2 \in \{c,n\})$, $E(Y(1,1) - Y(0,1)|T_1 = c, T_2 \in \{c,a\})$.

4.4 Direct and Indirect Effects (Causal Mechanisms)

As a further extension that is related to dynamic treatment effects, consider the problem of disentangling the total impact of a treatment into a direct effect and an indirect effect that operates via an intermediate variable (or so-called mediator) which also affects the outcome. That is, the interest lies in disentangling a treatment effect into various causal mechanisms. This may provide a better understanding of why specific treatments are effective or ineffective. As an example, consider the health effect of college attendance (D_1), which likely affects the employment state (D_2) which also influences the health outcome. Disentangling the direct effect of college attendance and its indirect effect operating via employment shows whether the health impact of college attendance is only driven by its impact on labor market participation, or also by other ("direct") channels such as college peers-induced changes of health behavior.

To formally define the effects of interest, let $D_2(d_1)$ denote the potential state of the second treatment as a function of the first. The standard notation for potential outcomes defined in terms of D_1 can then be easily linked to the notation appropriate to analysing causal mechanisms, namely: $Y(d_1) = Y(d_1, D_2(d_1))$, which makes explicit that D_1 might affect Y either directly or indirectly through its effect on D_2. Therefore, the LATE of the first treatment among compliers in the first treatment period can be expressed as

$$\Delta_{c1} = E\left(Y(1) - Y(0)|T_1 = c\right] = E[Y(1, D_2(1)) - Y(0, D_2(0))|T_1 = c), \tag{35}$$

and comprises both the direct and indirect effect of D_1 on Y.

The direct effect, denoted by $\theta_{c1}(d_1)$, is obtained by shutting down the indirect causal mechanism through fixing D_2 at its potential value under a particular d_1, while exogenously varying the first treatment D_1:

$$\theta_{c1}(d_1) = E\left(Y(1, D_2(d_1)) - Y(0, D_2(d_1))|T_1 = c\right), \qquad \text{for } d_1 \in \{0,1\}. \tag{36}$$

The indirect effect among compliers, denoted by $\delta_{c1}(d_1)$, corresponds to the mean difference in outcomes when exogenously shifting D_2 to its potential values for $d_1 = 1$ and $d_1 = 0$, while keeping the first treatment fixed at $D_1 = d_1$:

$$\delta_{c1}(d_1) = E\left(Y(d_1, D_2(1)) - Y(d_1, D_2(0))|T_1 = c\right), \qquad \text{for } d_1 \in \{0,1\}. \tag{37}$$

The LATE is the sum of the direct and indirect effects defined upon opposite states of d_1.This can be seen from adding and subtracting either $Y(0, D_2(1))$ or $Y(1, D_2(0))$ in (35):

$$\begin{aligned}\Delta_{c1} &= E\left(Y(1,D_2(1)) - Y(0,D_2(1))|T_1 = c\right) + E\left(Y(0,D_2(1)) - Y(0,D_2(0))|T_1 = c\right) = \theta_{c1}(1) + \delta_{c1}(0)\\ &= E\left(Y(1,D_2(0)) - Y(0,D_2(0))|T_1 = c\right) + E\left(Y(1,D_2(1)) - Y(1,D_2(0))|T_1 = c\right) = \theta_{c1}(0) + \delta_{c1}(1).\end{aligned}$$

The notation $\theta_{c1}(1), \theta_{c1}(0), \delta_{c1}(1), \delta_{c1}(0)$ implies that effects may be heterogenous with respect to d_1. This permits interaction effects between D_1 and D_2 on Y. In our health example, $\theta_{c1}(1)$ and $\theta_{c1}(0)$ are the direct effects of college attendance among first period compliers if their labor market states were set to their potential values with and without going to college.

Yamamoto (2013) shows identification of (36) and (37) based on an instrument for D_1 and a combined MAR and LI-type assumption (cf. Section 4.1) with respect to D_2: $Y(d_1, d_2) \perp D_2(d_1')|Z, T_1 = c, X$, for $d_1, d_1' \in \{0,1\}$. This allows controlling for the endogeneity of D_2 without a second IV. While D_1 and its instrument are both assumed to be binary, D_2 may be multivalued. Frölich and Huber (2017) base identification on two distinct instruments Z_1 and Z_2 for D_1 and D_2. While Z_1 and D_1 are assumed to the binary, the authors consider various sets of assumptions that yield (36) and (37) under continuous Z_2 and D_2, continuous Z_2 and discrete D_2, and discrete Z_2 and continuous D_2. Furthermore, they also discuss identification of the so-called controlled direct effect. Controlled direct effects are defined as the effect of D_1 when D_2 is fixed at a particular value d_2 for every complier (rather than its potential value $D_2(d_1)$) and also fit into the dynamic treatment effects framework: $E(Y(1,d_2) - Y(0,d_2)|T_1 = c)$.

5 Violations and Relaxations of the IV Assumptions

Sections 5.1 and 5.2 discuss violations of Assumption 1 and Assumption 2 (while maintaining Assumption 3) and under which relaxations effects on specific populations are nevertheless identified.

5.1 Violation of the Exclusion Restriction

We analyze the Wald estimand under violations of the exclusion restriction, i.e. part (ii) in Assumption 1, while maintaining instrument independence (part (i) of Assumption 1), monotonicity (Assumption 2), and a non-zero first stage (Assumption 3). As an example, consider the effect of military service on later life income when using a draft lottery as supposed IV, see for instance Angrist (1990). Proper implementation of the lottery satisfies part (i) of Assumption 1. However, part (ii) is violated if the lottery induces draft avoiding behavior that itself has an effect on income, thus constituting an alternative causal mechanism through which the IV affects the outcome. Enrollment to college represents such a mechanism if students are entitled to deferments that delay military conscription, see Card and Lemieux (2001).

We first consider a scenario in which the exclusion restriction is violated for noncompliers. Angrist, Imbens, and Rubin (1996) show that the Wald estimand equals the LATE plus a bias term given by

$$\frac{E(Y(1,D(1)) - Y(0,D(0))}{E(D(1) - D(0))} - E(Y(1,D(1)) - Y(0,D(0))|T = c) = E(H|T \neq c) \cdot \frac{1-\pi_c}{\pi_c},$$

where $H = Y(1,d) - Y(0,d)$ denotes the causal effect of Z on Y. Second, assume there is not only a direct effect of Z on Y for noncompliers but also for compliers. In addition, suppose that $Y(1,0) - Y(0,0) = Y(1,1) - Y(0,1)$ for all compliers, which allows expressing the causal effect of Z on Y as H and the causal effect of D on Y as $G = Y(z,1) - Y(z,0)$. Using this additional notation, the IV estimand is $\frac{E(Y(1,D(1))-Y(0,D(0))}{E(D(1)-D(0))} = E(G|T = c) + \frac{E(H)}{\pi_c}$, see Angrist, Imbens, and Rubin (1996). The second term gives the bias relative to the LATE and is

$$E(H|T = c) + E(H|T \neq c) \cdot \frac{1-\pi_c}{\pi_c}. \tag{38}$$

To interpret this result, consider the two components of the bias (38) separately. The first term originates from the direct effect of Z on Y for the compliers and does not depend on the compliance rate π_c. Thus, even under perfect compliance (that is if $\pi_c = 1$) this part of the bias would prevail whereas the second part would be zero. The second term equals the product of the direct effect of Z on Y for noncompliers and the odds of being a noncomplier. This implies that the sensitivity of the IV estimand to violations of the exclusion restriction depends on the strength of the instrument as measured by the size of the compliant population. The exclusion restriction is therefore crucial for point identification. In Section 6.2, we present alternative approaches to obtain bounds on the LATE under violations of the exclusion restriction.

5.2 Violation and Relaxations of Monotonicity

A violation of monotonicity (Assumption 2) implies the existence of defiers. As an example, consider the sex ratio of the first two siblings as IV for estimating the effect of having a third child on female labor supply as in Angrist and Evans (1998). Presuming that parents have a preference for mixed sex children, the idea is that having two children of the same sex, which is arguably randomly assigned by nature, induces some families to get a third child (compliers). However, monotonicity fails if some parents have a preference for at least two children of the same sex and choose to have a third child if the first two are of mixed sex. That such defiers may exist is for instance corroborated by Lee (2008), who finds that South Korean parents with one son and one daughter are more likely to continue childbearing than parents with two sons.

With defiers, the equality in (15) does not hold, see Angrist, Imbens, and Rubin (1996), but becomes

$$\frac{E(Y|Z=1)-E(Y|Z=0)}{E(D|Z=1)-E(D|Z=0)}=\Delta_c-\frac{\pi_d\cdot(\Delta_c-\Delta_d)}{\pi_c-\pi_d}=\frac{\pi_c\cdot\Delta_c-\pi_d\cdot\Delta_d}{\pi_c-\pi_d},$$

where $\Delta_d = E[Y(1)-Y(0)|T=d]$ is the LATE on defiers. Only in the special case that $\Delta_c = \Delta_d$ (homogenous LATEs on compliers and defiers) the LATE on compliers is identified.

When not imposing such a strong effect homogeneity assumption, the LATE is generally not obtained under a violation of Assumption 2. This does, however, not mean that nothing can be said about the LATE at all. Small and Tan (2007) and Small et al. (2017) show that the sign of Δ_c (as well as that of the ATE) is still identified if Assumption 2 is replaced by a weaker stochastic monotonicity condition, implying that $\Pr(T=c|A) \geq P(T=d|A)$ conditional on some latent stratum A. Stochastic monotonicity is, for instance, satisfied when A equals the potential outcomes such that $\Pr(T=c|Y(1),Y(0)) \geq P(T=d|Y(1),Y(0))$ (or $\Pr(T=c|U) \geq P(T=d|U)$ when assuming model (1)). That is, given any pair of potential outcome values under treatment and non-treatment, there must exist at least as many compliers as defiers. The same kind of assumption has also been considered in DiNardo and Lee (2011).

Treatment effects can be point identified even under specific relaxations of Assumption 2. Klein (2010) considers a nuisance term in the treatment equation that is unrelated with the potential outcomes and other unobserved factors affecting D (V in model (1)) and entails random departures from monotonicity such that some subjects defy. He discusses the conditions under which bias approximations for the identification of the LATE and MTE are obtained. Secondly, de Chaisemartin (2016) shows that the Wald estimand identifies the LATE among a subpopulation of compliers, which he denotes as "comvivors," if the following assumption holds:

Assumption 9

(Compliers-defiers). *There exists a subpopulation of compliers, denoted by* $T=cd$, *which satisfies* $\pi_{cd}=\pi_d$ *and* $E[Y(1)-Y(0)|T=cd]=\Delta_{cd}=\Delta_d$.

Assumption 9 states that some proportion of the total of compliers is equal to the defiers in terms of average effects and population size. Assumption 1, Assumption 3, and Assumption 9 identify the LATE on the remaining compliers not necessarily resembling the defiers, the so-called comvivors, which are defined as $T=cv:c$ without cd. This is the effect among those compliers who outnumber the compliers who resemble the defiers:

$$\frac{E(Y|Z=1)-E(Y|Z=0)}{E(D|Z=1)-E(D|Z=0)}=E(Y(1)-Y(0)|T=cv)=\Delta_{cv}.$$

Assumption 9 may appear abstract, but de Chaisemartin (2016) discusses several restrictions that imply this assumption and are easier to interpret. One possible condition is that compliers always outnumber defiers with the same treatment effect: $\Pr(T=c|Y(1)-Y(0)) \geq \Pr(T=d|Y(1)-Y(0))$, which is implied by, but weaker than the stochastic monotonicity assumption of Small and Tan (2007) and Small et al. (2017) discussed before. A second sufficient condition is that the LATEs of defiers and compliers have the same sign and that the ratio of the LATEs is not "too" large: Either $\mathrm{sgn}\Delta_d = \mathrm{sgn}\Delta_c \neq 0$ and $\Delta_d/\Delta_c \leq \pi_c/\pi_d$ or $\Delta_d=\Delta_c=0$. de Chaisemartin (2016) gives several empirical examples in which Assumption 2 appears unrealistic but Assumption 9 is arguably likely satisfied. Consider the evaluation of employment effects of disability insurance when using average allowance rates of randomly assigned examiners as instrument (e.g. Maestas, Mullen, and Strand 2013), or the effects of incarceration when using average sentencing rates of randomly assigned judges as an instrument (e.g. Aizer and Doyle 2013).

As a further strategy, Dahl, Huber, and Mellace (2016) replace Assumption 2 by a weaker local monotonicity condition given particular values of either marginal potential outcome distribution:

Assumption 10

(Local monotonicity). *Either* $\Pr(T=d|Y(d)=y(d))=0$ *or* $\Pr(T=c|Y(d)=y(d))=0$ *for* $d\in\{0,1\}$ *and all y(d) in the support of Y(d).*

Assumption 10 implies that conditional on a specific value of the potential outcome under treatment or non-treatment, either defiers or compliers do not exist. Thus, compliers and defiers are required to "inhabit" non-overlapping regions of the marginal potential outcome distributions.[5] Then, Δ_c is identified over all y satisfying $f(y,D=1|Z=1)-f(y,D=1|Z=0)>0$ and $f(y,D=0|Z=0)-f(y,D=0|Z=1)>0$ (see (12) and (13)), while Δ_d is over all y with $f(y,D=1|Z=1)-f(y,D=1|Z=0)<0$ and $f(y,D=0|Z=0)-f(y,D=0|Z=1)<0$. In most applications, a non-overlapping support in the potential outcomes of compliers and defiers appears unrealistic. However, local monotonicity might be combined with stochastic monotonicity-type assumptions to form conditions that seem more plausible.

6 Testing, Sensitivity Checks, and Bounds

Section 6.1 discusses approaches for testing the IV assumptions. Section 6.2 outlines sensitivity checks and bounds for the case that one is not willing to maintain Assumption 1–Assumption 3.

6.1 Testing the LATE Assumptions

Under Assumption 1 and Assumption 2, (12) and (13) not only permit evaluating local treatment effects, but also provide testable implications of the identifying assumptions. Namely, $f(y, D = 1|Z = 1) - f(y, D = 1|Z = 0) = f(y(1), T = c)$ and $f(y, D = 0|Z = 0) - f(y, D = 0|Z = 1) = f(y(0), T = c)$ imply for all y in the support of Y that

$$f(y, D = 1|Z = 1) \geq f(y, D = 1|Z = 0), \quad f(y, D = 0|Z = 0) \geq f(y, D = 0|Z = 1). \tag{39}$$

Otherwise, the joint densities of the compliers would be negative. Therefore, if one or both of the weak inequalities in (39) are violated, at least one of Assumption 1 and Assumption 2 is violated. These constraints were first derived by Balke and Pearl (1997) for binary outcomes, while Heckman and Vytlacil (2005) formulated them in terms of continuous outcomes. Note that the testable implications (39) remain unchanged when easing Assumption 2 to stochastic monotonicity of the form $\Pr(T = c|Y(d)) \geq \Pr(T = d|Y(d))$ for $d \in \{1, 0\}$, see Mourifié and Wan (2017).

For testing, (39) could be verified at each value y in the support of Y. However, if the outcome is of rich support (e.g. continuous), finite sample power may be higher when partitioning the support into a finite number of subsets. The testable constraints then are

$$\Pr(Y \in A, D = 1|Z = 1) \geq \Pr(Y \in A, D = 1|Z = 0), \Pr(Y \in A, D = 0|Z = 0) \geq \Pr(Y \in A, D = 0|Z = 1), \tag{40}$$

where A denotes a subset of the support of Y. Kitagawa (2015) proposes a test based on resampling a variance-weighted two sample Kolmogorov-Smirnov-type statistic on the supremum of $\Pr(Y \in A, D = 1|Z = 0) - \Pr(Y \in A, D = 1|Z = 1)$ and $\Pr(Y \in A, D = 0|Z = 1) - \Pr(Y \in A, D = 0|Z = 0)$ across multiple subsets A. The method can also be used for testing conditional on observed covariates, if the latter are binned into subsets of the support in a similar way as the outcomes. As an alternative approach, Mourifié and Wan (2017) show that a modified version of (40) making use of conditional moment inequality constraints fits the intersection bounds framework of Chernozhukov, Lee, and Rosen (2013). For binary outcomes, Machado, Shaikh, and Vytlacil (2018) propose a procedure that both verifies the constraints and the sign of the average treatment effect on the entire population. The authors also consider monotonicity of the outcome in the treatment as additional or alternative assumption to monotonicity of the treatment in the IV for testing.

As an alternative set of testable constraints, Huber and Mellace (2015) show that the LATE assumptions imply the following restrictions related to the mean potential outcomes (i) of the always takers under treatment and (ii) of the never takers under non-treatment:

$$\begin{aligned} &E(Y|D = 1, Z = 1, Y \leq y_q) \leq E(Y|D = 1, Z = 0) \leq E(Y|D = 1, Z = 1, Y \geq y_{1-q}), \\ &E(Y|D = 0, Z = 0, Y \leq y_r) \leq E(Y|D = 0, Z = 1) \leq E(Y|D = 0, Z = 0, Y \geq y_{1-r}). \end{aligned} \tag{41}$$

Under Assumption 1 and Assumption 2, $q = \Pr(D = 1|Z = 0)/\Pr(D = 1|Z = 1)$ gives the share of always takers among those with $D = 1$ and $Z = 1$, i.e. in the mixed population of compliers and always takers, and y_q is the qth quantile of Y given $D = 1$ and $Z = 1$. $r = 1 - (\Pr(D = 1|Z = 1) - \Pr(D = 1|Z = 0))/\Pr(D = 0|Z = 0)$ corresponds to the share of never takers among those with $D = 0$ and $Z = 0$, and y_r is the rth quantile of Y given $D = 0$ and $Z = 0$. Considering the first line of (41), the intuition of the test is as follows: $E(Y|D = 1, Z = 0)$ point identifies the mean potential outcome of the always takers under treatment, as any subject with $D = 1$, $Z = 0$ must be an always taker in the absence of defiers. Furthermore, the mean potential outcomes of the always takers are bounded by the averages in the upper and lower outcome proportions with $D = 1$ and $Z = 1$ that correspond to the share of the always takers in the mixed population: $E(Y|D = 1, Z = 1, Y \leq y_q)$, $E(Y|D = 1, Z = 1, Y \geq y_{1-q})$. $E(Y|D = 1, Z = 0)$ must lie within the latter bounds, otherwise the identifying assumptions are necessarily violated. An analogous result applies to the mean potential outcome of never takers under non-treatment.

It needs to be pointed out that any of the tests discussed so far check for necessary, albeit not sufficient conditions. That is, the methods are inconsistent in the sense there may exist counterfactual distributions which satisfy the testable restrictions, but violate Assumption 1 and Assumption 2. Sharma (2016) offers an extension to merely testing (40) by determining the likelihood that the LATE assumptions hold when the testable constraints are satisfied. To this end, the method defines classes of valid causal models satisfying the identifying assumptions as well as as invalid models and compares their marginal likelihood in the observed data.

Several further tests that are not based on constraints (40) have been proposed. Slichter (2014) suggests testing conditional IV validity (Assumption 4) by finding covariate values $X = x$ for which Z has no first stage and checking whether Z is associated with Y despite the absence of a first stage. For multivalued treatments, see Section 2.4, Angrist and Imbens (1995) argue that Assumption 2 implies that the cdfs of D given $Z = 1$ and $Z = 0$ do not cross, $\Pr(D \geq j|Z = 1) \geq \Pr(D \geq j|Z = 0)$, for $D, j \in \{0, 1,...,J\}$, which may be verified in the data. Fiorini and Stevens (2014) point out that testing this condition can have power against violations of both Assumption 2 and the independence of Z and $D(1)$, $D(0)$, which is part of Assumption 1. In the presence of both a binary and a continuous IV, Dzemski and Sarnetzki (2014) suggest an overidentification test. Finally, if outcome variables are observed in periods prior to IV and treatment assignment, placebo tests based on estimating the effect of Z on pre-instrument outcomes permit checking the plausibility of Assumption 1.

6.2 Sensitivity Checks and Bounds

If IV validity appears questionable, one may consider sensitivity checks on the robustness of the LATE under violations of the IV assumptions or the derivation of upper and lower bounds on the effect under weaker assumptions. Huber (2014), for instance, proposes sensitivity checks under the non-satisfaction of the IV exclusion restriction (inherent in Assumption 1) or Assumption 2. Under a presumed violation of the exclusion restriction while maintaining the random assignment of Z and Assumption 2 and Assumption 3, the LATE can be shown to correspond to

$$\frac{E(Y|Z = 1) - E(Y|Z = 0) - \Pr(D = 1|Z = 0) \cdot \gamma_a - \Pr(D = 0|Z = 1) \cdot \gamma_n}{\Pr(D = 1|Z = 1) - \Pr(D = 1|Z = 0)} - \gamma_c. \tag{42}$$

γ_c, γ_a, and γ_n denote the average direct effects of Z to the mean potential outcomes of the compliers, always takers, and never takers. Jones (2015) derives a related result under the assumption that the direct effect on the never takers (γ_n) is equal to zero. Under a homogeneous direct mean effect across types, implying that $\gamma_a = \gamma_n = \gamma_c = \gamma$, (42) simplifies to $(E(Y|Z = 1) - E(Y|Z = 0) - \gamma)/(\Pr(D = 1|Z = 1) - \Pr(D = 1|Z = 0))$. Inference can therefore be conducted if the researcher has a plausible prior about the range of values $\gamma_c, \gamma_a, \gamma_n$, or γ might take, see also Conley, Hansen, and Rossi (2012). The approach suggested by Slichter (2014) based on his IV validity test (see Section 6.1) may be used for determining such values in sensitivity checks.

Under a violation of Assumption 2 while maintaining Assumption 1 and Assumption 3, Huber (2014) shows that the mean potential outcomes of the compliers correspond to:

$$\begin{aligned} E(Y(1)|T = c) &= \frac{\Pr(D = 1|Z = 1) \cdot E(Y|D = 1, Z = 1)}{\rho_a(\Pr(D = 1|Z = 0) - \pi_d) + \Pr(D = 1|Z = 1) - \Pr(D = 1|Z = 0) + \pi_d}, \\ E(Y(0)|T = c) &= \frac{\Pr(D = 0|Z = 0) \cdot E(Y|D = 0, Z = 0)}{\rho_n(\Pr(D = 0|Z = 1) - \pi_d) + \Pr(D = 1|Z = 1) - \Pr(D = 1|Z = 0) + \pi_d}, \end{aligned} \tag{43}$$

where $\rho_a = E(Y(1)|T = a)/E(Y(1)|T = c)$ and $\rho_n = E(Y(0)|T = n)/E(Y(0)|T = c)$ are the ratios of mean potential outcomes (i) of always takers and compliers under treatment and (ii) of never takers and compliers under non-treatment. Considering various combinations of ρ_a, ρ_n, and π_d allows for investigating the sensitivity of the LATE to violations of monotonicity.

If plausible values for the aforementioned tuning parameters appear hard to justify, a bounds analysis may seem more credible, at the likely cost of a larger set of potential LATE values. Flores and Flores-Lagunes (2013) maintain Assumption 2 and the random assignment of Z, but assume a violation of the exclusion restriction. They impose restrictions on the order of specific mean potential outcomes (i) across treatment or instrument states within particular types and (ii) across types to narrow the bounds. Mealli and Pacini (2013) tighten the bounds using an auxiliary variable for which the exclusion restriction holds (contrary to the outcome), e.g. a covariate measured prior to randomization, and which is associated with the outcome and/or the compliance type. Richardson and Robins (2010) maintain Assumption 1, but assume a violation of Assumption 2 and derive bounds for the LATEs of various compliance types when the outcome is binary. Under mean independence of Z and the potential outcomes/treatments, Huber, Laffers, and Mellace (2017) bound the LATEs on several subpopulations when monotonicity is violated, with and without invoking a particular order in the mean potential outcomes across types.

7 External Validity of the LATE

Section 7.1 discusses checks for the external validity of the LATE based on observables. Section 7.2 presents conditions for extrapolating the LATE to the ATE and tests thereof. Section 7.3 considers the partial identification of the ATE based on the IV assumptions.

7.1 Comparability in Terms of Observables

Comparing compliers and the total population in terms of observed characteristics may be useful for judging the plausibility of the LATE being (close to) externally valid. Angrist and Fernández-Val (2010) consider $\Pr(X = x|T = c)/\Pr(X = x)$, the relative likelihood of covariate values $X = x$ among compliers compared to the entire population, which is identified under Assumption 1–Assumption 3 by the ratio of the first stage given $X = x$ to the overall first stage:

$$\frac{\Pr(X = x|T = c)}{\Pr(X = x)} = \frac{\Pr(T = c|X = x)}{\Pr(T = c)} = \frac{E(D|Z = 1, X = x) - E(D|Z = 0, X = x)}{E(D|Z = 1) - E(D|Z = 0)}.$$

Furthermore, under the conditional IV assumptions (Assumption 4–Assumption 7), the mean or other distributional features of the covariates among compliers can be obtained by using the κ-weighting function of Abadie (2003) provided in Section 3.5: $E(X|T = c) = \frac{E(\kappa \cdot X)}{E(\kappa)}$. While such checks may provide important insights about the representativeness of compliers in terms of X, the caveat remains that nothing can be said about unobserved characteristics.

7.2 Conditions for Extrapolation and Testing

This section discusses assumptions that allow extrapolating from the LATE to the ATE. Angrist (2004) distinguishes two restrictions under which the LATE is directly externally valid, i.e. corresponds to the ATE. Under the first restriction, there is no selection in the sense that mean potential outcomes under either treatment state are constant across types. Under the second restriction, selection is quite specific such that the levels of the mean potential outcomes differ across types, but the mean effects do not, i.e. are homogeneous. Note that under the first restriction, both (i) the Wald estimand and (ii) $E(Y|D = 1) - E(Y|D = 0) = E(Y(1)|T \in \{c, a\}) - E(Y(0)|T \in \{c, n\})$ identify the ATE because $E(Y(d)|T) = E(Y(d))$. Under the second restriction, only (i) but not (ii) yields the effect, as $E(Y(d)|T) \neq E(Y(d))$. Angrist (2004), Brinch, Mogstad, and Wiswall (2017), and Huber (2013) consider tests for the external validity of the LATE under the first restriction based on differences in mean potential outcomes across types. To see this, consider the following regression, which is fully nonparametric as D and Z are binary:

$$\begin{aligned}
E(Y|D, Z) &= \beta_0 + \beta_D D + \beta_Z Z + \beta_{DZ} DZ, \text{ where} \\
\beta_Z &= E(Y|D = 0, Z = 1) - E(Y|D = 0, Z = 0) = E(Y(0)|T = n) - E(Y(0)|T \in \{c, n\}), \\
\beta_{DZ} &= E(Y|D = 1, Z = 1) - E(Y|D = 0, Z = 1) - \{E(Y|D = 1, Z = 0) - E(Y|D = 0, Z = 0)\} \\
&= E(Y(1)|T \in \{c, a\}) - E(Y(0)|T = n) - \{E(Y(1)|T = a) - E(Y(0)|T \in \{c, n\})\}.
\end{aligned}$$

β_Z captures selection driven by the difference in mean potential outcomes of compliers and never takers under non-treatment (or differences in U under model (1)). β_{DZ} reflects both selection and treatment effect heterogeneity across types.

If $\beta_Z = 0$, β_{DZ} simplifies to $E(Y(1)|T \in \{c, a\}) - E(Y(1)|T = a)$, and a difference may be due to selection or differential treatment effects across always takers and compliers. Therefore, jointly testing for β_Z and β_{DZ} has in general non-trivial power to detect heterogeneity in mean potential outcomes across types, either driven by selection or treatment effect heterogeneity, which generally implies that the LATEs differ across types, too. However, not all potential violations can be tested, as the potential outcomes of never takers under treatment and always takers under non-treatment are never observed. Strictly speaking, $\beta_Z = \beta_{DZ} = 0$ is therefore not sufficient for external validity under the first restriction of Angrist (2004). Furthermore, it is not necessary either, because the LATE can theoretically be homogeneous across groups even if potential outcomes differ (second restriction of Angrist (2004)). However, under the assumption that differences in mean potential outcomes either occur across all or across no types (i.e. either $E(Y(d)|T) = E(Y(d))$ or $E(Y(d)|T = t) \neq E(Y(d)|T = t')$ for any $t \neq t'$ and $t, t' \in \{c, a, n\}$), β_Z suffices for detecting selection. Conditional on $\beta_Z = 0$, β_{DZ} in this case exclusively detects effect heterogeneity across types. For this reason, it appears worthwhile testing $\beta_Z = \beta_{DZ} = 0$, which can be easily implemented by means of an F-test.[6] In the case of one-sided noncompliance ($\Pr(D(0) = 1) = 0$), only $E(Y|D = 0, Z = 1) - E(Y|D = 0, Z = 0) = 0$ is testable.

If Assumption 4–Assumption 7 rather than Assumption 1–Assumption 3 are satisfied, the same type of test may be conducted conditional on X, see de Luna and Johansson (2014) and Black et al. (2015). In this case, testing can also be framed as a check for the conditional mean independence of the treatment and the potential outcomes given observed covariates: $E(Y(d)|D, X) = E(Y(d)|X)$, which would imply that $E(Y|D = 1, X) - E(Y|D = 0, X)$ yielded the causal effect $E(Y(1) - Y(0)|X)$. Donald, Hsu, and Lieli (2014b) suggest an alternative approach to test this condition based on a comparison of treatment effects under one-sided noncompliance, ruling out always takers and defiers. As in the latter case the LATT ($\Delta_{c,\,D=1}$) corresponds to the ATT ($\Delta_{D=1}$), Donald, Hsu, and Lieli (2014b) construct a Durbin-Wu-Hausmann-type test based on the z-statistic for the difference of the respective estimates $\hat{\Delta}_{c,\,D=1}$ and $\hat{\Delta}_{D=1}$. These estimates are obtained by the sample analogue of (29) and the approach suggested in Hirano, Imbens, and Ridder (2003). While the satisfaction of conditional mean independence of the treatment implies that identification and extrapolation can be achieved without an instrument, an instrument is required to obtain an overidentifying restriction and being able to construct a test for $E(Y(d)|D, X) = E(Y(d)|X)$. Note that if the latter assumption holds, not only the LATE on compliers, ATE, and ATT are identified, but also the LATEs on never and always takers, as discussed in Frölich and Lechner (2015).

If the strong assumption of effect homogeneity across types is not satisfied, the LATE may nevertheless permit extrapolation to the ATE even under a binary instrument if particular parametric assumptions hold, see Brinch, Mogstad, and Wiswall (2017). They assume the MTE $\Delta(\bar{V} = p(z))$ of Section 2.3 to be linear in $p(Z) = \Pr(D = 1|Z)$ by imposing linearity on $E(Y(0)|p(Z))$ and $E(Y(1)|p(Z))$, see also Restriction 3 in Angrist (2004). Brinch, Mogstad, and Wiswall (2017) furthermore demonstrate that polynomial (rather than linear) MTE functions are identified if the MTE is additively separable in X and unobservable factors and if X satisfies specific support conditions.

As an alternative source of extrapolation, Angrist and Fernández-Val (2010) and Aronow and Carnegie (2013) consider homogeneity of the ATE given X across types, a conditional version of effect homogeneity under the second restriction of Angrist (2004):

Assumption 11

(Conditional effect homogeneity). $E(Y(1) - Y(0)|T, X) = E(Y(1) - Y(0)|X)$

Assumption 11 implies that heterogeneity in average effects across types is solely due to X, such that $\Delta_c(x) = E(Y(1) - Y(0)|X = x) = \Delta(x)$. Therefore, the ATE, denoted as Δ, is obtained by $\int \Delta_c(x) dF_X(x) = E(\Delta_c(x))$ if the conditional LATE assumptions (Assumption 4–Assumption 7) are satisfied.[7] More generally, the ATE on some population selected by the binary indicator S (e.g. $S = D$ for the treated and $S = 1 - D$ for the nontreated) corresponds to

$$\Delta_{S=1} = \int \Delta_c(x) dF_{X|S=1}(x) = \int \Delta_c(x) \frac{\Pr(S = 1|X)}{\Pr(S = 1)} dF_X(x) = E\left[\Delta_c(x) \frac{\Pr(S = 1|X)}{\Pr(S = 1)}\right].$$

It is noteworthy that conditional effect homogeneity rules out that effect heterogeneity is driven by unobserved gains, which importantly restricts the source of treatment effect heterogeneity and is not consistent with the Roy (1951) model. Angrist and Fernández-Val (2010) demonstrate that Assumption 11 is testable (conditional on the satisfaction of Assumption 4–Assumption 7) if more than one instrument is available. Denote by $\Delta_c^W(x)$ and $\Delta_c^Z(x)$ the conditional LATEs based on two different instruments W and Z. It must hold that

$$\Delta_{S=1} = \int \Delta_c^W(x) \frac{\Pr(S = 1|X)}{\Pr(S = 1)} dF_X(x) = \int \Delta_c^Z(x) \frac{\Pr(S = 1|X)}{\Pr(S = 1)} dF_X(x).$$

See also Heckman, Schmierer, and Urzua (2010) for testing approaches in the context of the MTE framework that verify conditional effect homogeneity based on multiple instruments.[8]

Another extrapolation strategy is based on the rank invariance assumption discussed in Section 8.2. Wüthrich (2018) and Vuong and Xu (2017) show that the counterfactual mappings,

$$P_{01|T=t} = Q_{Y(0)|T=t}\left(F_{Y(1)|T=t}(y)\right) \text{ and } P_{10|T=t} = Q_{Y(1)|T=t}\left(F_{Y(0)|T=t}(y)\right),$$

which relate each individual outcome to its counterfactual, do not depend on the type T under rank invariance. Hence, one can use $P_{01|T=c}(y)$ and $P_{10|T=c}(y)$, which are identified under Assumption 1–Assumption 3, for imputing the distributions of $Y(1)$ for never takers and of $Y(0)$ for always takers. This is the intuition underlying the IV quantile regression model of Section 8.2.

7.3 Partial Identification of the ATE

Even if point identification of the ATE fails because the LATE estimates are not externally valid, the identifying power of the IV assumptions may be used to partially identify the ATE and other parameters (such as the ATT) not discussed here. Balke and Pearl (1997) (for binary outcomes) as well as Heckman and Vytlacil (2001a) and Kitagawa (2009) (for more general outcomes) derive bounds on the ATE under Assumption 1–Assumption 3. They also provide the interesting result that these bounds coincide with the bounds of Manski (1990), who merely assumes $E(Y(d)|Z=1) = E(Y(d)|Z=0)$ for $d \in \{1, 0\}$. Shaikh and Vytlacil (2011) sharpen the bounds on the ATE in the binary outcome case under the assumption that the treatment effect is either weakly positive or weakly negative for all individuals (while the direction is a priori not restricted). Cheng and Small (2006) extend the results for binary outcomes to the case that the treatment can take three values under particular forms of (one-sided) noncompliance.

Under mean independence of Z and the potential outcomes/treatments and Assumption 2, Huber, Laffers, and Mellace (2017) bound the ATE when assuming a particular order in the mean potential outcomes across types. Chen, Flores, and Flores-Lagunes (2017) consider such restrictions in addition to Assumptions 1 and 2, too, but also invoke a specific order of mean potential outcomes across treatment states within specific types. Furthermore, see Chiburis (2010) and references therein for the derivation of semiparametric (rather than nonparametric) bounds on the ATE under the IV assumptions. Kowalski (2016) considers the MTE framework and assumes the marginal outcomes under treatment and non-treatment, $E(Y(1)|V=v)$ and $E(Y(0)|V=v)$, to be monotonic in the unobserved term of treatment model (1) to bound the ATE. Angrist (2004) offers a sensitivity check for the ATE based on particular proportionality conditions across the mean potential outcomes of various types. Finally, Mogstad, Santos, and Torgovitsky (2017) develop a framework for obtaining identified sets on the ATE and other parameters by exploiting the fact that the IV estimand and many other parameters of interest can be expressed as a weighted average of the MTE.

8 Relationship to Other Instrumental Variable Approaches

We discuss the relationship between the LATE setup and two popular IV models: the linear IV model with covariates (Section 8.1) and the IV quantile regression model (Section 8.2).

8.1 Linear IV Models

Linear IV models such as $Y = X'\gamma + \beta D + \varepsilon$ play a central role in applied research, with X now comprising a constant and the vector of covariates, γ and β denoting coefficients, and ε representing the error term. If treatment effects are homogeneous across individuals, then β is constant and corresponds to the population ATE. The latter can thus be consistently estimated using classical estimators such as TSLS or limited information maximum likelihood (LIML). In most applications it appears implausible that treatment effects are homogeneous and thus unrelated to observables or unobservables. It is therefore important to understand which parameters linear IV methods estimate when treatment effects are in fact heterogeneous.

To formalize the analysis, we follow Angrist and Imbens (1995) and Angrist and Pischke (2009) and consider the TSLS estimand with fully saturated first and second stage equations

$$D = \pi_X + \pi_{1X} Z + u, \quad Y = \alpha_X + \beta D + \varepsilon.$$

π_X and α_X denote saturated models for the covariates and π_{1X} denotes separate first-stage effects of Z for every value of X. Under the assumptions of the LATE framework with covariates (Assumption 4–Assumption 7) it can be shown that

$$\beta = E\left(\omega(X) \cdot \Delta_c(X)\right), \quad \text{where} \quad \omega(X) = \frac{Var\left(E(D|X,Z)|X\right)}{E\left(Var\left(E(D|X,Z)|X\right)\right)}.$$

That is, TSLS with a fully saturated first stage and a second stage which is saturated in the covariates converges to a weighted average of covariate-specific LATEs. The weights are proportional to the average conditional variance of the population first-stage fitted value $E(D \mid X, Z)$.

Kolesar (2013) generalizes the analysis in Angrist and Imbens (1995) by characterizing the estimands of general two-step estimators (such as TSLS) and minimum distance estimators (such as LIML) under the LATE framework. His analysis shows that while the probability limit of TSLS can be expressed as a convex combination of LATEs as in the special case discussed before, that of LIML and related estimators may be outside the convex hull of LATEs. Such estimands may thus not correspond to a causal effect under effect heterogeneity.

8.2 IV Quantile Regression

The instrumental variable quantile regression (IVQR) model introduced by Chernozhukov and Hansen (2005) provides an alternative framework for evaluating heterogeneous treatment effects with IVs. In contrast to the LATE framework, the IVQR model does not impose a monotonicity assumption in the selection equation. Instead, it relies on rank invariance in the outcome equation, a restriction on the evolution of individual ranks across treatment states.[9] By rank invariance, the IVQR model identifies population level treatment effects. This is in sharp contrast to the LATE framework under which treatment effects are only identified among compliers. However, rank invariance substantially restricts treatment effect heterogeneity and may therefore be implausible in many applications. As noted by Heckman and Vytlacil (2007), rank invariance rules out scenarios in which agents self-select based on their individual effects and does not allow effect heterogeneity as generated by the generalized Roy model.

To formalize rank invariance, note that by the Skorohod representation of random variables, the potential outcome $Y(d)$ can be related to its quantile function $Q_{Y(d)}(\tau)$ as follows:

$$Y(d) = Q_{Y(d)}(U(d)), \text{ where } U(d) \sim \text{Uniform}(0,1). \tag{44}$$

If the potential outcomes are continuous random variables, $Q_{Y(d)}(\cdot)$ is strictly increasing and the disturbance $U(d)$ determines the individual position or rank in the distribution of $Y(d)$. We therefore refer to $U(d)$ as "rank." With this notation at hand, one can formally define rank invariance as $U(1) = U(0)$. Under rank invariance and IV validity, the population QTE, $\Delta(\tau) = Q_{Y(1)}(\tau) - Q_{Y(0)}(\tau)$, is identified by the following conditional moment restriction:

$$\Pr\left(Y \leq Q_{Y(D)}(\tau)|Z\right) = \tau. \tag{45}$$

On the surface, the IVQR and the LQTE model do not seem to be connected since they rely on different, non-nested assumptions and identify treatment effects for different populations. Despite these differences, Wüthrich (2018) shows that

$$\Delta(\tau) = \Delta_c\left(F_{Y(0)|T=c}\left(Q_{Y(0)}(\tau)\right)\right) = \Delta_c\left(F_{Y(1)|T=c}\left(Q_{Y(1)}(\tau)\right)\right), \tag{46}$$

where $Q_{Y(1)}$ and $Q_{Y(0)}$ are the inverses of $F_{Y(1)}$ and $F_{Y(0)}$ which are defined as

$$\begin{aligned} F_{Y(1)}(y) &= \pi_a F_{Y(1)|T=a}(y) + \pi_c F_{Y(1)|T=c}(y) + \pi_n F_{Y(0)|T=n}\left(Q_{Y(0)|T=c}\left(F_{Y(1)|T=c}(y)\right)\right), \\ F_{Y(0)}(y) &= \pi_n F_{Y(0)|T=n}(y) + \pi_c F_{Y(0)|T=c}(y) + \pi_a F_{Y(1)|T=a}\left(Q_{Y(1)|T=c}\left(F_{Y(0)|T=c}(y)\right)\right). \end{aligned}$$

Equation (46) shows that the IVQR QTE estimand at quantile level τ corresponds to the LQTE at τ', where τ is generally not equal to τ'. The difference between the estimands is determined by two factors: (i) differences between the potential outcome distributions of the untreated compliers and never takers as well as differences between the potential outcome distributions of the treated compliers and always takers, and (ii) the relative size of the three subpopulations.

The results in Wüthrich (2018) confirm that if treatment effect heterogeneity is unrestricted, all the information about treatment effects has to come from the compliant subpopulation, i.e. the only subpopulation for which potential outcomes are identified under both treatment states. Furthermore, the results demonstrate that IVQR achieves identification by extrapolating from the compliers to the overall population. This motivates the use of IVQR for extrapolation in the LATE framework, see Section 7.2.

9 Conclusion

This paper provides a survey on the methodological advancements in the evaluation of local average treatment effects based on instruments. We first review the classical framework going back to the seminal contributions of

Imbens and Angrist (1994) and Angrist, Imbens, and Rubin (1996), which have been very influential in applied empirical research. We then proceed by summarizing and synthesizing important methodological extensions, for example distributional and quantile treatment effects, multivalued or multiple treatments and instruments, identification and estimation in the presence of observed covariates, attrition and measurement error, testing and relaxations of identifying assumptions, conditions for external validity, and the relationship to other IV approaches. We thereby complement more introductory reviews that focus on implementation and applications such as Imbens (2014) and the textbook discussions in Angrist and Pischke (2009, 2015).

Acknowledgement

We have benefitted from comments by Martin Eckhoff Andresen, Michael Knaus, Anna Solovyeva, Svitlana Tyahlo, and an anonymous referee.

Notes

[1] IV-based identification can also be obtained in structural models that are different to (1) and (2). This concerns for instance the relation of Z and D, see for instance Hernan and Robins (2006): $Y = \phi(D,U), \quad D = \eta(V,U), \quad Z = \kappa(V), \quad U \perp V$, where $\phi(\cdot)$, $\eta(\cdot)$, $\kappa(\cdot)$ are unknown functions. Here, D is not affected by Z, but the variables are correlated through V so that Z predicts D. As V and U are independent, Assumption 1 holds. See also Chalak and White (2011), who exhaustively discuss structural relations under which Z is a valid IV.

[2] Conventionally, TSLS has been used for linear IV models with homogeneous effects. TSLS then converges to $Cov(Y, Z)/Cov(D, Z)$ for scalar instrument and treatment variables, under the conditions that Z has a zero covariance with the error term of Y and a non-zero covariance with Z. The latter two conditions are implied by Assumption 1-Assumption 3, while effect homogeneity is not. Therefore, TSLS estimates the average effect on the compliers in our nonparametric context, rather than a constant effect as in the linear context. Furthermore, $Cov(Y,Z)/Cov(D,Z) = \frac{Cov(Y,Z)}{Var(Z)} \Big/ \frac{Cov(D,Z)}{Var(Z)}$, which equals the Wald estimand $\frac{E(Y|Z=1)-E(Y|Z=0)}{E(D|Z=1)-E(D|Z=0)}$ for a binary Z.

[3] One might therefore be tempted to binarize the multivalued treatment based on a specific threshold in its support, e.g. using a binary indicator for completing high school as treatment rather than years of schooling. However, this generally violates the IV exclusion restriction, see Marshall (2016) and Andresen and Huber (2018).

[4] Monotonicity requires that changing Z from 0 to 1 (2) affects treatment choice 1 (2) vs. 0, but not 2 (1).

[5] As an example, reconsider the quarter of birth instrument for education and redshirting (postponement of school entry) as source of defiers, which more frequently occurs among families with a high socio-economic status, see Bedard and Dhuey (2006) and Aliprantis (2012). Assumption 10 would be satisfied if the socioeconomic status determined both defiance and the potential outcomes in a deterministic way, e.g. if children coming from defying families with a high socio-economic status had higher potential earnings than children of complying ones.

[6] The following approaches are asymptotically equivalent to such an F-test: Testing (i) $\frac{E(Y|Z=1)-E(Y|Z=0)}{E(D|Z=1)-E(D|Z=0)} = E(Y|D=1) - E(Y|D=0)$ as in the classical Hausman (1978) test, (ii) $\frac{E(Y|Z=1)-E(Y|Z=0)}{E(D|Z=1)-E(D|Z=0)} = E(Y|D=1, Z=0) - E(Y|D=0, Z=1)$ as suggested in Angrist (2004), and (iii) $E(Y|D=1, Z=1) = E(Y|D=1, Z=0)$, $E(Y|D=0, Z=1) = E(Y|D=0, Z=0)$ as considered by Bertanha and Imbens (2015).

[7] Under Assumption 11, Assumption 5 might even be relaxed, e.g. to stochastic monotonicity given X.

[8] Kédagni and Mourifié (2016) discuss yet another approach for testing the external validity of the LATE with a single instrument, however, when relying on a different set of restrictions than the standard LATE assumptions. Notably, they assume (i) that the instrument and the potential outcomes are uncorrelated rather than fully independent and (ii) that the potential outcomes have bounded support.

[9] Chernozhukov and Hansen (2005) show that rank invariance can be somewhat relaxed to rank similarity that allows for random deviations from the expected rank.

References

Abadie, A. 2002. “Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models.” *Journal of the American Statistical Association* 97: 284–292.

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

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

Aizer, A., and J. J. Doyle. 2013. “Juvenile Incarceration, Human Capital and Future Crime: Evidence from Randomly-Assigned Judges.” Technical report, NBER.

Aliprantis, D. 2012. “Redshirting, Compulsory Schooling Laws, and Educational Attainment.” *Journal of Educational and Behavioral Statistics* 37: 316–338.

Andresen, M. E., and M. Huber. 2018. “Instrument-based Estimation with Binarized Treatments: Issues and Tests for the Exclusion Restriction.” SES Working Paper 492, University of Fribourg.

Angrist, J. 1990. “Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records.” *American Economic Review* 80: 313–336.

Angrist, J., and W. Evans. 1998. “Children and their Parents Labor Supply: Evidence from Exogeneous Variation in Family Size.” *American Economic Review* 88: 450–477.

Angrist, J., and I. Fernández-Val. 2010. "Extrapolate-ing: External Validity and Overidentification in the Late Framework." NBER working paper 16566.

Angrist, J., and G. W. Imbens. 1995. "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity." *Journal of American Statistical Association* 90: 431–442.

Angrist, J., G. W. Imbens, and D. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of American Statistical Association* 91: 444–472 (with discussion).

Angrist, J., and A. Krueger. 1991. "Does Compulsory School Attendance Affect Schooling and Earnings?" *Quarterly Journal of Economics* 106: 979–1014.

Angrist, J. D. 2004. "Treatment Effect Heterogeneity in Theory and Practice." *The Economic Journal* 114: C52–C83.

Angrist, J. D., and J.-S. Pischke. 2009. *Mostly Harmless Econometrics: An Epiricist's Companion*. Princeton University Press.

Angrist, J. D., and J.-S. Pischke. 2015. *Mastering 'Metrics: The Path from Cause to Effect*, Princeton: Princeton University Press.

Aronow, P. M., and A. Carnegie. 2013. "Beyond Late: Estimation of the Average Treatment Effect with an Instrumental Variable." *Political Analysis* 21: 492–506.

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

Barua, R., and K. Lang. 2009. "School Entry, Educational Attainment, and Quarter of Birth: A Cautionary Tale of Late." NBER Working Paper 15236.

Battistin, E., M. De Nadai, and B. Sianesi. 2014. "Misreported Schooling, Multiple Measures and Returns to Educational Qualifications." *Journal of Econometrics* 181: 136–150.

Bedard, K., and E. Dhuey. 2006. "The Persistence of Early Childhood Maturity: International Evidence of Long-Run Age Effects." *The Quarterly Journal of Economics* 121: 1437–1472.

Behaghel, L., B. Crépon, and M. Gurgand. 2013. "Robustness of the Encouragement Design in a Two-Treatment Randomized Control Trial." IZA Discussion Paper No 7447.

Belloni, A., V. Chernozhukov, I. Fernández-Val, and C. Hansen. 2017. "Program Evaluation and Causal Inference with High-Dimensional Data." *Econometrica* 85: 233–298.

Bertanha, M., and G. Imbens. 2015. "External Validity in Fuzzy Regression Discontinuity Designs." NBER working paper 20773.

Black, D. A., J. Joo, R. J. LaLonde, J. A. Smith, and E. J. Taylor. 2015. "Simple Tests for Selection Bias: Learning More from Instrumental Variables." IZA Discussion Paper No 9346.

Blackwell, M. 2015. "Identification and Estimation of Joint Treatment Effects with Instrumental Variables." working paper, Department of Government, Harvard University.

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

Bound, J., D. A. Jaeger, and R. M. Baker. 1995. "Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogeneous Explanatory Variable is Weak." *Journal of the American Statistical Association* 90: 443–450.

Brinch, C. N., M. Mogstad, and M. Wiswall. 2017. "Beyond Late with a Discrete Instrument." *Journal of Political Economy* 125: 985–1039.

Buckles, K. S., and D. M. Hungerman. 2013. "Season of Birth and Later Outcomes: Old Questions, New Answers." *Review of Economics and Statistics* 95: 711–724.

Card, D. 1995. "Using Geographic Variation in College Proximity to Estimate the Return to Schooling." In *Aspects of Labor Market Behaviour: Essays in Honour of John Vanderkamp*, edited by L. Christofides, E. Grant, and R. Swidinsky, 201–222. Toronto: University of Toronto Press.

Card, D., and T. Lemieux. 2001. "Going to College to Avoid the Draft: The Unintended Legacy of the Vietnam War." *The American Economic Review* 91: 97–102.

Carneiro, P., J. J. Heckman, and E. J. Vytlacil. 2011. "Estimating Marginal Returns to Education." *American Economic Review* 101: 2754–2781.

Carneiro, P., and S. Lee. 2009. "Estimating Distributions of Potential Outcomes Using Local Instrumental Variables With an Application to Changes in College Enrollment and Wage Inequality." *Journal of Econometrics* 149 (2): 191–208.

Chalak, K. 2016 . "Instrumental Variables Methods with Heterogeneity and Mismeasured Instruments." *Econometric Theory* 33: 1– 36.

Chalak, K., and H. White. 2011. "An Extended Class of Instrumental Variables for the Estimation of Causal Effects." *Canadian Journal of Economics* 44: 1–51.

Chen, X., and C. A. Flores. 2015. "Bounds on Treatment Effects in the Presence of Sample Selection and Noncompliance: The Wage Effects of Job Corps." *Journal of Business & Economic Statistics* 33: 523–540.

Chen, X., C. A. Flores, and A. Flores-Lagunes. 2017. "Going Beyond Late: Bounding Average Treatment Effects of Job Corps Training." *Journal of Human Resources*. DOI: 10.3368/jhr.53.4.1015.7483R1.

Cheng, J., and D. S. Small. 2006. "Bounds on Causal Effects in Three-Arm Trials with Non-compliance." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68: 815–836.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. 2017. "Double/Debiased Machine Learning for Treatment and Structural Parameters." *Econometrics Journal* 21: C1–C68.

Chernozhukov, V., and C. Hansen. 2005. "An IV Model of Quantile Treatment Effects." *Econometrica* 73: 245–261.

Chernozhukov, V., S. Lee, and A. Rosen. 2013. "Intersection Bounds: Estimation and Inference." *Econometrica* 81: 667–737.

Chiburis, R. C. 2010. "Semiparametric Bounds on Treatment Effects." *Journal of Econometrics* 159: 267–275.

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

Cornelissen, T., C. Dustmann, A. Raute, and S. Uta. 2016. "From Late to MTE: Alternative Methods for the Evaluation of Policy Interventions." IZA DP No. 10056.

Dahl, C. M., M. Huber, and G. Mellace. 2016. "It's Never Too Late. A New Look at the Identification of Local Average Treatment Effects with or Without Defiers." working paper, University of Southern Denmark, Dept. of Economics.

de Chaisemartin, C. 2016. "Tolerating Defiance? Identification of Treatment Effects Without Monotonicity." working paper, University of Warwick.

de Luna, X., and P. Johansson. 2014. "Testing for the Unconfoundedness Assumption Using an Instrumental Assumption." *Journal of Causal Inference* 2: 187–199.

Deaton, A. S. 2010. "Instruments, Randomization, and Learning about Development." *Journal of Economic Literature* 48: 424–455.

DiNardo, J., and D. S. Lee. 2011. "Program Evaluation and Research Designs," In *Handbook of Labor Economics*, edited by Orley Ashenfelter, and David Card, Vol. 4, 463–536. New York: Elsevier.

DiTraglia, F., and C. Garcia-Jimeno. 2016. "On Mis-Measured Binary Regressors: New Results and Some Comments on the Literature." working paper, University of Pennsylvania.

Donald, S. G., Y.-C. Hsu, and R. P. Lieli. 2014a. "Inverse Probability Weighted Estimation of Local Average Treatment Effects: A Higher Order MSE Expansion." *Statistics and Probability Letters* 95: 132–138.

Donald, S. G., Y.-C. Hsu, and R. P. Lieli. 2014b. "Testing the Unconfoundedness Assumption via Inverse Probability Weighted Estimators of (L)ATT." *Journal of Business & Economic Statistics* 32 (3): 395–415.

Dzemski, A., and F. Sarnetzki. 2014. "Overidentification Test in a Nonparametric Treatment Model with Unobserved Heterogeneity." mimeo, University of Mannheim.

Fiorini, M., and K. Stevens. 2014. "Monotonicity in IV and fuzzy RD designs - A Guide to Practice." mimeo, University of Sydney.

Flores, C. A., and A. Flores-Lagunes. 2013. "Partial Identification of Local Average Treatment Effects With an Invalid Instrument." *Journal of Business & Economic Statistics* 31: 534–545.

Frangakis, C., and D. Rubin. 1999. "Addressing Complications of Intention-to-Treat Analysis in the Combined Presence of All-or-None Treatment-Noncompliance and Subsequent Missing Outcomes." *Biometrika* 86: 365–379.

Fricke, H., M. Frölich, M. Huber, and M. Lechner. 2015. "Endogeneity and Non-Response Bias in Treatment Evaluation: Nonparametric Identification of Causal Effects by Instruments." IZA Discussion Paper No 9428.

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

Frölich, M., and M. Huber. 2014. "Treatment Evaluation with Multiple Outcome Periods Under Endogeneity and Attrition." *Journal of the American Statistical Association* 109: 1697–1711.

Frölich, M., and M. Huber. 2017. "Direct and Indirect Treatment Effects - Causal Chains and Mediation Analysis with Instrumental Variables." *Journal of the Royal Statistical Society Series B* 79: 1645–1666.

Frölich, M., and M. Lechner. 2015. "Combining Matching and Nonparametric Instrumental Variable Estimation: Theory and an Application to the Evaluation of Active Labour Market Policies." *Journal of Applied Econometrics* 30: 718–738.

Frölich, M., and B. Melly. 2013a. "Identification of Treatment Effects on the Treated with One-Sided Non-Compliance." *Econometric Reviews* 32: 384–414.

Frölich, M., and B. Melly. 2013b. "Unconditional Quantile Treatment Effects Under Endogeneity." *Journal of Business & Economic Statistics* 31: 346–357.

Frumento, P., F. Mealli, B. Pacini, and D. B. Rubin. 2012. "Evaluating the Effect of Training on Wages in the Presence of Noncompliance, Nonemployment, and Missing Outcome Data." *Journal of the American Statistical Association* 107: 450–466.

Hausman, J. A. 1978. "Specification Tests in Econometrics." *Econometrica* 46: 1251–1271.

Heckman, J. J. 1997. "Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations." *The Journal of Human Resources* 32: 441–462.

Heckman, J. J., and R. Pinto. 2018. "Unordered Monotonicity." *Econometrica* 86: 1–35.

Heckman, J. J., D. Schmierer, and S. Urzua. 2010. "Testing the Correlated Random Coefficient Model." *Journal of Econometrics* 158: 177–203.

Heckman, J. J., and S. Urzúa. 2010. "Comparing IV with Structural Models: What Simple IV Can and Cannot Identify." *Journal of Econometrics* 156: 27–37.

Heckman, J. J., and E. Vytlacil. 1999. "Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects." *Proceedings National Academic Sciences USA*, Economic Sciences 96, 4730–4734.

Heckman, J. J., and E. Vytlacil. 2001a. "Instrumental Variables, Selection Models, and Tight Bounds on the Average Treatment Effects." In *Econometric Evaluation of Labour Market Policies*, edited by M. Lechner, M. Pfeiffer, 1–15. New York: Center for European Economic Research.

Heckman, J. J., and E. Vytlacil. 2001b. "Local Instrumental Variables." In *Nonlinear Statistical Inference: Essays in Honor of Takeshi Amemiya*, edited by C. Hsiao, K. Morimune, J. Powell. Cambridge: Cambridge University Press.

Heckman, J. J., and E. Vytlacil. 2005. "Structural Equations, Treatment Effects, and Econometric Policy Evaluation 1." *Econometrica* 73: 669–738.

Heckman, J. J., and E. J. Vytlacil. 2007. "Chapter 71 Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast their Effects in New Environments." In *Part B of Handbook of Econometrics*, edited by James J. Heckman, Edward E. Leamer, Vol. 6, 4875–5143. North-Holland: Elsevier.

Hernan, M. A., and J. M. Robins. 2006. "Instruments for Causal Inference. An Epidemiologist's Dream?" *Epidemiology* 17: 360–372.

Hirano, K., G. W. 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.

Hsu, Y.-C., T.-C. Lai, and R. P. Lieli. *Estimation and Inference for Distribution Functions and Quantile Functions in Endogenous Treatment Effect Models* 2015 Working Paper, Central European University.

Huber, M. 2013. "A Simple Test for the Ignorability of Non-Compliance in Experiments." *Economics Letters* 120: 389–391.

Huber, M. 2014. Sensitivity Checks for the Local Average Treatment Effect." *Economics Letters* 123: 220–223.

Huber, M., L. Laffers, and G. Mellace. 2017. "Sharp IV Bounds on Average Treatment Effects on the Treated and Other Populations Under Endogeneity and Noncompliance." *Journal of Applied Econometrics* 32: 56–79.

Huber, M., and G. Mellace. 2015. "Testing Instrument Validity for LATE Identification Based on Inequality Moment Constraints." *Review of Economics and Statistics* 97: 398–411.

Hull, P. 2015. "Isolateing: Identifying Counterfactual-Specific Treatment Effects with Cross-Stratum Comparisons." working paper, MIT Department of Economics.

Imbens, G. W. 2004. “Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review.” *The Review of Economics and Statistics* 86: 4–29.

Imbens, G. W. 2010. “Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009).” *Journal of Economic Literature* 48 (2): 399–423.

Imbens, G. W. 2014. “Instrumental Variables: An Econometrician’s Perspective.” IZA Discussion Paper No. 8048.

Imbens, G. W., and J. Angrist. 1994. “Identification and Estimation of Local Average Treatment Effects.” *Econometrica* 62: 467–475.

Imbens, G. W., and D. Rubin. 1997. “Estimating Outcome Distributions for Compliers in Instrumental Variables Models.” *Review of Economic Studies* 64: 555–574.

Imbens, G. W., and J. M. Wooldridge. 2009. “Recent Developments in the Econometrics of Program Evaluation.” *Journal of Economic Literature* 47: 5–86.

Jones, D. 2015. “The Economics of Exclusion Restrictions in IV Models.” NBER working paper 21391, Cambridge, MA.

Kédagni, D., and I. Mourifié. 2016. “Empirical Content of the IV Zero-Covariance Assumption: Testability, Partial Identification.” working paper, University of Toronto.

Kirkeboen, L., E. Leuven, and M. Mogstad. 2016. “Fields of Study, Earnings, and Self-Selection.” *Quarterly Journal of Economics* 131: 1057–1111.

Kitagawa, T. 2009. “Identification Region of the Potential Outcome Distribution Under Instrument Independence.” CeMMAP working paper 30/09.

Kitagawa, T. 2015. “A Test for Instrument Validity.” *Econometrica* 83: 2043–2063.

Klein, T. J. 2010. “Heterogeneous Treatment Effects: Instrumental Variables Without Monotonicity?” *Journal of Econometrics* 155: 99–116.

Kolesar, M. 2013. *Estimation in an Instrumental Variable Model with Treatment Effect Heterogeneity*, Working Paper, Princeton University.

Kowalski, A. E. 2016. “Doing More When You’re Running Late: Applying Marginal Treatment Effect Methods to Examine Treatment Effect Heterogeneity in Experiments.” working paper, Yale University.

Lee, J. 2008. “Sibling Size and Investment in Children’s Education: An Asian Instrument.” *Journal of Population Economics* 21: 855–875.

Lee, S., and B. Salanie. 2015. “Identifying Effects of Multivalued Treatments.” cemmap working paper CWP72/15.

Little, R., and D. Rubin. 1987. *Statistical Analysis with Missing Data*. New York: Wiley.

Machado, C., A. Shaikh, and E. Vytlacil. 2018. *Instrumental Variables, and the Sign of the Average Treatment Effect*, Working Paper, University of Chicago.

Maestas, N., K. J. Mullen, and A. Strand. 2013. “Does Disability Insurance Receipt Discourage Work? Using Examiner Assignment to Estimate Causal Effects of SSDI Receipt.” *The American Economic Review* 103: 1797–1829.

Manski, C. F. 1990. “Nonparametric Bounds on Treatment Effects.” *American Economic Review*, 319–323. Papers and Proceedings 80.

Marshall, J. 2016. “Coarsening Bias: How Coarse Treatment Measurement Upwardly Biases Instrumental Variable Estimates.” *Political Analysis* 24: 157–171.

Mealli, F., G. Imbens, S. Ferro, and A. Biggeri. 2004. “Analyzing a Randomized Trial on Breast Self-examination with Noncompliance and Missing Outcomes.” *Biostatistics* 5: 207–222.

Mealli, F., and B. Pacini. 2013. “Using Secondary Outcomes and Covariates to Sharpen Inference in Instrumental Variable Settings.” *Journal of the American Statistical Association* 108: 1120–1131.

Melly, Blaise, and Kaspar Wüthrich. 2018. “Local Quantile Treatment Effects,” In *Handbook of Quantile Regression*, edited by Roger Koenker, Victor Chernozhukov, Xiuming He, and Limin Peng, 145–164. Chapman and Hall/CRC.

Miquel, R. 2002. “Identification of Dynamic Treatment Effects by Instrumental Variables.” University of St. Gallen Economics Discussion Paper Series 2002–2011.

Mogstad, M., A. Santos, and A. Torgovitsky. 2017. *Using Instrumental Variables for Inference about Policy Relevant Treatment Effects*, NBER Working Paper No. 23568.

Mourifié, I., and Y. Wan. 2017. “Testing Late Assumptions.” *The Review of Economics and Statistics* 99: 305–313.

Richardson, T. S., and J. M. Robins. 2010. “Analysis of the Binary Instrumental Variable Model.” In *Heuristics, Probability and Causality: A Tribute to Judea Pearl*, edited by R. Dechter, H. Geffner, and J. Y. Halpern, 415–440. London, UK: College Publications.

Rosenbaum, P., and D. Rubin. 1983. “The Central Role of the Propensity Score in Observational Studies for Causal Effects.” *Biometrika* 70: 41–55.

Roy, A. 1951. “Some Thoughts on the Distribution of Earnings.” *Oxford Economic Papers* 3: 135–146.

Rubin, D. B. 1974. “Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies.” *Journal of Educational Psychology* 66: 688–701.

Rubin, D. B. 1976. “Inference and Missing Data.” *Biometrika* 63: 581–592.

Shaikh, A., and E. Vytlacil. 2011. “Partial Identification in Triangular Systems of Equations with Binary Dependent Variables.” *Econometrica* 79: 949–955.

Sharma, A. 2016. “Necessary and Probably Sufficient Test for Finding Valid Instrumental Variables.” working paper, Microsoft Research, New York.

Slichter, D. 2014. *Testing Instrument Validity and Identification with Invalid Instruments*. Mimeo: University of Rochester.

Small, D. S., and Z. Tan. 2007. “A Stochastic Monotonicity Assumption for the Instrumental Variables Method.” Technical report, Department of Statistics, Wharton School, University of Pennsylvania.

Small, D. S., Z. Tan, R. R. R. S. A. Lorch, and M. A. Brookhart. 2017. “Instrumental Variable Estimation with a Stochastic Monotonicity Assumption.” *Statistical Science* 32: 561–579.

Tan, Z. 2006. “Regression and Weighting Methods for Causal Inference Using Instrumental Variables.” *Journal of the American Statistical Association* 101: 1607–1618.

Ura, T. 2016. “Heterogeneous Treatment Effects with Mismeasured Endogenous Treatment.” working paper, Duke University.

Uysal, S. D. 2011. *Doubly Robust IV Estimation of the Local Average Treatment Effects*. Mimeo: University of Konstanz.

Vuong, Q., and H. Xu. 2017. “Counterfactual Mapping and Individual Treatment Effects in Nonseparable Models with Binary Endogeneity.” *Quantitative Economics* 8 (2): 589–610.

Vytlacil, E. 2002. “Independence, Monotonicity, and Latent Index Models: An Equivalence Result.” *Econometrica* 70: 331–341.

Wüthrich, Kaspar 2018. “A Comparison of Two Quantile Models with Endogeneity.” Journal of Business and Economic Statistics, Accepted for publication.

Yamamoto, T. 2013. “Identification and Estimation of Causal Mediation Effects with Treatment Noncompliance.” unpublished manuscript, MIT Department of Political Science.

Yanagi, T. 2017. “Inference on Local Average Treatment Effects for Misclassified Treatment.” working paper, Hitotsubashi University, Tokyo.

Yu, P. 2014. *Marginal Quantile Treatment Effect and Counterfactual Analysis*, Working Paper, The University of Hong Kong.

Zhang, J., D. Rubin, and F. Mealli. 2009. “Likelihood-Based Analysis of Causal Effects of Job-Training Programs Using Principal Stratification.” *Journal of the American Statistical Association* 104: 166–176.